5-2009

Modeling Potential Native Plant Species Distributions in Rich County, Utah

Kathryn A. Peterson
Utah State University

MODELING POTENTIAL NATIVE PLANT SPECIES DISTRIBUTIONS IN RICH COUNTY, UTAH

by

Kathryn A. Peterson

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Range Science

Approved:

_____	_____
R. Douglas Ramsey Major Professor	Eugene W. Schupp Committee Member
_____	_____
Janis L. Boettinger Committee Member	Byron Burnham Dean of Graduate Students

UTAH STATE UNIVERSITY
Logan, Utah

2008

ABSTRACT

Modeling Potential Native Plant Species Distributions in Rich County, Utah

by

Kathryn A. Peterson, Master of Science

Utah State University, 2008

Major Professor: R. Douglas Ramsey
Department: Wildland Resources

Georeferenced field data were used to develop logistic regression models of the geographic distribution of 38 frequently common plant species throughout Rich County, Utah, to assist in the future correlation of Natural Resources Conservation Service Ecological Site Descriptions to soil map units. Field data were collected primarily during the summer of 2007, and augmented with previously existing data collected in 2001 and 2006. Several abiotic parameters and Landsat Thematic Mapper imagery were used to stratify the study area into sampling units prior to the 2007 field season.

Models were initially evaluated using an independent dataset extracted from data collected by the Bureau of Land Management and by another research project conducted in Rich County by Utah State University. By using this independent dataset, model accuracy statistics widely varied across individual species, but the average model sensitivity (modeling a species as common where it was common in the independent dataset) was 0.626, and the average overall correct classification rate was 0.683. Because of concerns pertaining to the appropriateness of the independent dataset for evaluation,

models were also evaluated using an internal cross-validation procedure. Model accuracy statistics computed by this procedure averaged 0.734 for sensitivity and 0.813 for overall correct classification rate. There was less variability in accuracy statistics across species using the internal cross-validation procedure.

Despite concerns with the independent dataset, we wanted to determine if models would be improved, based on internal cross-validation accuracy statistics, by adding these data to the original training data. Results indicated that the original training data, collected with this modeling effort in mind, were better for choosing model parameters, but sometimes model coefficients were better when computed using the combined dataset.

(107 pages)

ACKNOWLEDGMENTS

I would like to thank my major professor, Doug Ramsey, and committee members Janis Boettinger and Gene Schupp. I will never forget the opportunities and support I have received from Doug and others at Utah State University. This project was funded by the Natural Resource Conservation Service (NRCS) and the Utah Agricultural Experiment Station. I am very thankful for the financial support I received from these institutions, and the technical support I and others received from the NRCS's State Range Conservationist for Utah, Shane Green.

Special thanks to Alexander Hernandez, who was always there to provide good advice. Others that deserve to be recognized for their assistance, both technical and emotional, include John Lowry, Leila Shultz, and Lisa Langs-Stoner.

My former colleagues at the Oregon Bureau of Land Management and NRCS need to be acknowledged as well – Ed Horn, Charlie Tackman, Tom Clark, Larry Thomas, and many others. They taught me a lot about rangeland ecology, soils, and ecosystem function; many of the things that I learned from these professionals were applied to this thesis.

Also, thank you to my dad and brothers who influenced me in so many ways, especially in encouraging my love of learning and science.

I would especially like to thank Jeff Brown, my husband, for his encouragement, excellent suggestions, and patience. Without him, this endeavor would have been much more difficult, if not impossible.

Kathryn A. Peterson

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Ecological Site Descriptions and Soil Map Units

An Ecological Site is defined as "a distinctive kind of land with specific physical characteristics that differs from other kinds of land in its ability to produce a distinctive kind and amount of vegetation" (U.S. Department of Agriculture, National Soil Survey Handbook, Part 622.07, 2007). Ecological Site Descriptions (ESDs) provide land managers with information that can be used to facilitate appropriate land use and management. In the western United States, understanding biotic community dynamics and potential land use impacts is particularly important for balancing needs for livestock production with the desire to maintain ecosystem function and biotic diversity.

The Natural Resources Conservation Service (NRCS), in cooperation with the Bureau of Land Management (BLM) and Bureau of Indian Affairs (BIA), is the agency that is responsible for the development (and revision when necessary) of ecological site descriptions (ESDs) (U.S. Department of Agriculture, National Range and Pasture Handbook, Part 600.0103, 2003). NRCS ESDs are usually correlated to soil series (named and described soil types) in conjunction with the development of soil surveys. By utilizing NRCS tools such as Web Soil Survey (available at: http://websoilsurvey.nrcs.usda.gov/app/), Soil Data Viewer (available at: http://soildataviewer.nrcs.usda.gov/default.aspx) or Soil Data Mart downloads (available at: http://soildatamart.nrcs.usda.gov/), managers can obtain spatial and tabular soils and ESD correlation information. ESDs can be obtained via the NRCS's Electronic Field Office Technical Guide (available at: http://www.nrcs.usda.gov/Technical/efotg/).

NRCS soil survey spatial data currently exists as polygon coverages composed of soil map units. Most surveys conducted in areas of less intensive land use are at a scale of 1:24,000 (Soil Survey Division Staff, 1993) or less. Because of scale limitations, soil map unit polygons may contain more than one soil series, and usually contain at least some small areas of soil inclusions. Series descriptions do not reflect the actual character of soils at all locations where it is found within polygons; it is simply a description of a "typical" pedon.

The concept that environmental factors and biotic communities are inseparably linked to soils was described by V.V. Dokuchaev in the 1880s (Buol et al., 2003). His ideas were popularized in the United States by *Factors of Soil Formation: A System of Quantitative Pedology* by Hans Jenny (1941). Jenny's model, S = *f*(cl, o, r, p, t, …), relates soil distribution (S) to climate (cl), organisms (o), relief (r), parent material (p), and time (t). Though this model has been revised since 1941 (Gerrard, 1981), it is still used as a conceptual framework.

Recently, McBratney et al. (2003) proposed a very Jenny-like model: Sc or Sp = *f*(s, c, o, r, p, a, n), where Sc = soil class, Sp = soil property, s = other soil properties at a point, c = climatic properties, o = organisms (floral, faunal, or human), r = relief or topography, p = parent material or lithology, a = age or time factor, and n = space or spatial position. This new model attempts to formally recognize spatial concepts and the idea that soil properties at specific locations are related to other soil properties at those same locations. The model also indicates that soil information produced from it can be either discreet (such as drainage classes), or continuous (such as horizon depth).

Because soil characteristics and other abiotic factors strongly influence native biotic communities, ESDs are based on abiotic attributes rather than biotic community characterizations. Particular combinations of abiotic factors result in the distinct vegetation communities seen across landscapes, assuming that natural disturbance regimes are acting upon native biotic communities which have not been greatly altered by post-European settlement human activity. Also, areas with similar suites of abiotic attributes should respond similarly to the same management activities. Using abiotic factors rather than more easily observed plant community characteristics to define an ESD is an ideal approach when one is interested in determining what type of native biotic community an area is capable of supporting when no native species are currently present.

Species vs. ESD Modeling

An ESD is developed from data collected at sampling locations that are typical of the type of community that the ESD defines. This usually requires the presence of communities relatively free of significant human impacts; however, there are no known relic communities (*sensu* Clements, 1928) in Rich County (N.E. West, Rangeland Ecologist, Retired, personal communication; S. Green, NRCS State Range Conservationist for Utah, personal communication). Although ESDs have been developed that describe potential biotic communities in the county, it is difficult to determine *where* they occur without more detailed information. Also, there is the possibility that ESDs do not exist for some of the major types of potential plant communities that occur within the county, or that some existing ESDs might need to be modified to reflect local potential biotic communities more accurately.

Plant Species Dominance and Stability

Plant community composition involves not only the identity and number of species, but abundance of each species as well. Species composition arises partly from deterministic processes linking habitat characteristics to species-specific niches and stochastic processes such as seed dispersal (Ozinga et al., 2005). Typically plant communities are composed of a small set of relatively abundant species mixed with a larger number of minor species (Hall, 1992; Walker et al., 1999). In this thesis, plant species are considered relatively abundant or common if they compose the greatest proportion of cover or biomass within the plant community. Generally, species that have proportionally more foliar cover also compose a greater proportion of the total biomass, although there are exceptions to this rule. At almost all of the sampling sites for this study there were fewer than four species within each life form (forbs, grasses, shrubs, or trees) that composed more than 1% of the foliar canopy cover; these were considered to be the common species.

At any particular site, the dominant or most common species are considered to be the best adapted species for the local suite of abiotic and biotic factors, and serve to maintain ecosystem function (Walker et al., 1999). Vegetation associations are the result of seed availability and environmental selection, and environments are principally determined by climate and soil altered by physiographic and biotic processes. Where climatic and physiographic changes are slow, continued migration and species interactions tend to produce relatively stable and static vegetation assemblages (Gleason, 1926). Predictive modeling of species distributions relies on the assumption that an equilibrium exists between biotic communities and abiotic factors (Guisan and Theurillat,

2000). This premise is necessarily restricted to limited temporal scales, and can not apply where communities are undergoing rapid succession or other type of change.

Approaches to Predictive Modeling and Accuracy Assessment

Predictive modeling of species distributions can be separated into two types: mechanistic and correlative (or empirical). Mechanistic models require detailed knowledge of species' resource needs (Robertson et al., 2003), and the ability to determine or model these resources accurately. Correlative models, on the other hand, are not expected to describe real cause-and-effect relationships between plant species response and model variables (Guisan and Zimmermann, 2000). Predictive vegetation models are generally correlative (Guisan and Zimmermann, 2000), and are often based on parameters that indirectly affect resource gradients. Results of work by Robertson et al. (2003) suggest that correlative vegetation prediction models may perform as well as mechanistic models.

A variety of statistical techniques have been used to infer species distributions from presence/absence, presence only, and abundance data (see Guisan and Zimmermann, 2000). Some of the more common techniques for presence/absence data include generalized linear modeling (GLM), generalized additive modeling (GAM), classification trees, and Bayesian approaches. The most straightforward to implement in a geographic information system (GIS) is GLM. Logistic regression is one type of GLM which provides estimated probabilities of occurrence on a 0- to-1 scale.

It is generally accepted that accuracy of a model should be assessed using an independent dataset (i.e. data not used to develop the model). Using the same data to

evaluate (or validate) the model as was used to develop it (resubstitution) tends to produce an overly optimistic accuracy estimate. The most robust estimate of model accuracy can be achieved using an independent dataset, particularly if that dataset was generated using a different sampling strategy than the data used for model building (Guisan and Zimmermann, 2000). For this project, two independent datasets were available for accuracy assessment. One was from fieldwork conducted by Utah State University for a wildlife (passerine) study, and another dataset collected for an ecological site inventory by Bureau of Land Management (BLM) rangeland specialists. If all available data are used to build the model, there are internal cross-validation techniques such as bootstrapping or *k*-fold partitioning that can be used to provide a more robust estimate of accuracy than resubstitution (Fielding and Bell, 1997; Guisan and Zimmermann, 2000).

Abiotic Factors Data

Many abiotic attribute datasets are either available online or can be computed from other data layers using GIS software. Several of these datasets consist of grids having specific spatial resolutions and thematic accuracies that affect the models utilizing them. For example, continuous climatic data have been developed using elevation-driven interpolations of climate station data (Guisan and Zimmermann, 2000). Standard climatic datasets include PRISM (Parameter-Elevation Regressions on Independent Slopes Model, available at: http://www.prism.oregonstate.edu/) and Daymet (available at: http://www.daymet.org/) precipitation and temperature data for the continental United States at 800 m^2 and 1 km^2 resolution, respectively.

Errors in the interpolation process and/or lack of sufficient weather station data introduce spatial uncertainty into these maps. Also, due to the distance between stations (sample locations), the resolution of map products do not provide microclimate information that affect plant communities at finer resolutions (Guisan and Zimmermann, 2000). For these reasons climatic maps are often replaced by or augmented with digital elevation models (DEM) which are available at finer resolutions. Because elevation grids are used as a primary model parameter for generating climate maps, the elevation grid itself and its derivatives (e.g. slope and aspect) are often used as surrogates for temperature and sometimes precipitation grids.

DEM data are available for the contiguous United States at 30 m^2 resolution from the U.S. Geological Survey's National Elevation Dataset (http://ned.usgs.gov/). These data have a vertical accuracy of ± 7-15 m (http://ned.usgs.gov/Ned/faq.asp). DEMs provide not only elevation data, but various other indirect gradients including slope, slope aspect, and slope curvature can be derived from them. Specific catchment area (Tarboton, 1997) and potential clear-sky solar radiation, both direct and diffuse (Kumar et al., 1997), can be modeled from DEMs as well. Although these factors may have no direct influence on the suitability of an area for plant establishment, they do affect microclimates – surface temperatures and water availability – that can affect plant distributions.

It is also possible to differentiate areas with different parent materials and/or soils from remotely-sensed imagery. Black-and-white aerial photography has been utilized by soil scientists to aid in the production of soil maps since the 1930s (Kornblau and Cipra, 1983). The use of remotely sensed imagery for mapping soils is limited by its ability to

only view surficial characteristics that are not obscured by vegetation cover (Grunwald and Lamsal, 2006). On the other hand, vegetation indices such as the Normalized Difference Vegetation Index (NDVI) can provide information to aid soil differentiation or classification, particularly when used in combination with DEM data (see Dobos et al., 2000). Band ratios and principal components analysis (PCA) have also been used to enhance or highlight multiband spectral reflectance characteristics to differentiate soil types (see van Deventer, 1992; Martínez-Rios and Monger, 2002).

The Rich County Soil Survey

The soil survey for Rich County, Utah, was completed in 1980; most of the field work was done in the 1970s (Campbell and Lacey, 1982). At that time, the predecessors to ESDs, Rangesite Descriptions, were being correlated to soil map units by the NRCS. Rangesite Descriptions were generally broader in scope and less detailed than ESDs. Subsequent to the development of ESDs and the migration of soil survey information into the National Soil Information System (NASIS), Rangesite Descriptions were 'translated' into the ESDs that were used to populate the NASIS database.

Recent fieldwork conducted in Rich County by the Bureau of Land Management (BLM) and other agencies have indicated that some of the ESDs currently correlated to soil map units in NASIS are not appropriate for this landscape. Several ESDs were developed in Major Land Resource Areas (MLRAs) (U.S. Department of Agriculture, Handbook 296, 2006) that do not intersect Rich County and contain inappropriate potential species compositions and/or productivity data. The review of ESD correlations in Rich County prompted the attempt to develop potential plant species distribution models.

In the published soil survey map unit descriptions (Campbell and Lacey, 1982) there is a very brief description of vegetation found within soil map units at the time of the survey. For most units there is an equally brief interpretation of the potential historic climax plant community. This is typical for soil surveys – they are not focused on vegetation, but on soils. Unfortunately, this information is not sufficient to make soil component-to-ESD correlations. One major limitation is the fact that the published soil survey does not differentiate between *Artemisia tridentata* Nutt. ssp. *tridentata* (basin big sagebrush), *A. tridentata* ssp. *vaseyana* (mountain big sagebrush), and *A. tridentata* ssp. *wyomingensis* (Wyoming big sagebrush). At the time of the survey, there was less interest in the differentiation between these subspecies of big sagebrush than there is currently. Differentiation between the subspecies of *A. tridentata* is necessary to make appropriate ESD correlations. Also, although species listed as present at the time of the soil survey provide some indication of the species that have the potential to occur on those soils, they do not include all of the potentially common species. Some interpretations for potential climax plant community also appear inappropriate – as when *Artemisia nova* (black sagebrush) is listed as an existing species, but "big sagebrush" is listed as a potential species. These two species, whose distributions are largely determined by soil characteristics (Zamora and Tueller, 1973), are unlikely to occupy the same types of sites.

The State of Rangelands in Rich County

Unfortunately the historic character of many of the sagebrush communities of the Intermountain West will never be known accurately because these areas were heavily grazed by livestock or came under cultivation soon after European settlement (Ellison,

1960). Studies have shown that grazing has an effect on the relative abundance and composition of species in plant communities (del-Val and Crawley, 2005). Improper livestock grazing by cattle in semi-arid rangelands has been shown to result in an increase in shrubs and/or juniper and a decrease in grass cover and diversity (Ellison, 1960).

Plant communities throughout Rich County have been altered to varying degrees from their natural or historic states. Existing plant species compositions may be completely different from their natural or potential states, or they may retain many species with altered species abundance patterns. This project attempts to calculate the probability that specific plant species will be common at any given location by extrapolating current species distribution data across the landscape. A basic assumption in this study is that common species at sample locations have the potential to be common at those sites and at other sites with the same suite of abiotic attributes.

Project Objectives

The goal of this research project was to develop species distribution maps to assist in the review and re-correlation of ESDs to soil map units in Rich County. A spatial modeling approach was taken due to the lack of representative field data in each of the 138 soil maps units to effectively re-correlate ESDs using a more conventional approach. Furthermore, budget restrictions and available time negated a comprehensive field campaign to achieve this goal. Therefore, the specific project objectives were to:

1. Stratify the Rich County landscape using abiotic factors that affect microclimate water availability and temperature to optimize field sampling within a limited time frame.
2. Spatially model the potential distribution of common plant species.

References

Buol, W.W., Southard, R.J., Graham, R.C., McDaniel, P.A., 2003. Soil Genesis and Classification, 5th ed. Iowa State Press, Blackwell Publishing Company. Ames, 494 pp..

Campbell, L.B., Lacey, C.A., 1982. Soil Survey of Rich County, Utah. Soil Conservation Service, U.S. Department of Agriculture, 273 pp.

Clements, F.E., 1928. Plant Succession and Indicators: A Definitive Edition of Plant Succession and Plant Indicators. H. W. Wilson Company, New York, 453 pp.

del-Val, E., Crawley, M.J., 2005. Are grazing increaser species better tolerators than decreasers? An experimental assessment of defoliation tolerance in eight British grassland species. Ecology 93:1005-1016.

Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., Helt, T., 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma 97:367-391.

Ellison, L., 1960. Influence of grazing on plant succession of rangelands. The Botanical Review. 26:1-78.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24(1), 38-49.

Gerrard, A.J., 1981. Soils and Landforms: An Integration of Geomorphology and Pedology. George Allen & Unwin (Publishers) Ltd, London, 219 pp.

Gleason, H.A., 1926. The individual concept of the plant association. In: McIntosh, R.P. (Ed.), Benchmark Papers in Ecology v. 6 – Phytosociology. Bull. Torrey Bot. Club. 53:7-26.

Grunwald, S., Lamsal S., 2006. The impact of emerging geographic information technology on soil-landscape modeling. Environmental soil-landscape modeling: geographic information technologies and pedometrics. In: Grunwald, S. (Ed.), CRC Press, Taylor & Francis Group, Boca Raton, Florida, 488 pp.

Guisan, A., Theurillat, J., 2000. Equilibrium modeling of alpine plant distribution: how far can we go? Phytocoenologia, 30(3-4):353-384.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135:147-186.

Hall, C.A.S., Stanford, J.A., Hauer, F.R., 1992. The distribution and abundance of organisms as a consequence of energy balances along multiple environmental gradients. Oikos, 65:377-390.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill. New York, 281pp.

Kornblau, M.L., Cipra, J.E., 1983. Investigation of digital Landsat data for mapping soils under range vegetation. Remote Sensing of Environment, 13:103-112.

Kumar, L., Skidmore, A.K., Knowles, E., 1997. Modelling topographic variation in solar radiation in a GIS environment. Intl. J. Geo. Info. Sci. 11(5): 475-497.

Martínez-Ríos, J.J., and Monger, H.C., 2002. Soil classification in arid lands with Thematic Mapper data. Terra, 20:89-100.

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma, 117:3-52.

Ozinga, W.A., Schaminée, J.H.J., Bekker, R.M., Bonn, S., Poschlod, P., Tackenberg, O., Bakker, J., van Groenendael, J.M., 2005. Predictability of plant species composition from environmental conditions is constrained by dispersal limitation. Oikos, 108:555-561.

Robertson, M.P., Peter, C.I., Villet, M.H., Ripley, B.S., 2003. Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques. Ecol. Model. 164:153-167.

Soil Survey Division Staff, 1993. Soil survey manual. Soil Conservation Service. U.S. Department of Agriculture Handbook 18.

Tarboton, D.G., 1997. A new method for the determination of flow directions and contributing areas in grid digital elevation models. Water Resour. Res. 33(2): 309-319.

U.S. Department of Agriculture, Natural Resources Conservation Service, 2003. National Range and Pasture Handbook, title 190-VI-NRPH, Revision 1. [Online] Available: http://www.glti.nrcs.usda.gov/technical/publications/nrph.html. Accessed 14 Jan 2007.

U.S. Department of Agriculture, Natural Resources Conservation Service, 2007. National Soil Survey Handbook, title 430-VI. [Online] Available: http://soils.usda.gov/technical/handbook/. Accessed 14 Jan 2007.

U.S. Department of Agriculture, 2006. Land Resource Regions and Major Land Resource Areas of the United States, the Caribbean, and the Pacific Basin, Handbook 296, [Online] Available: http://soils.usda.gov/survey/geography/mlra/index.html. Accessed 15 Jan 2007.

van Deventer, A.P., 1992. Evaluating the usefulness of Landsat Thematic Mapper data to determine soil properties, management practices and soil water content. Ph.D. Dissertation. The Ohio State University, Columbus, Ohio, 342 pp.

Walker, B., Kinzig, A., Langridge, J., 1999. Plant attribute diversity, resilience, and ecosystem function: The nature and significance of dominant and minor species. Ecosystems, 2:95-113.

Zamora, B., Tueller, P.T., 1973. *Artemisia arbuscula*, *A. longiloba*, and *A. nova* habitat types in northern Nevada. Great Basin Naturalist, 33(4): 225-242.

CHAPTER 2

USING GIS-DERIVED CLUSTERS OF ABIOTIC FACTORS TO GUIDE A LIMITED FIELD-SAMPLING EFFORT[1]

Summary

1. Field sampling over a large area with diverse terrain can present many challenges. Completely random sampling may not be feasible given limited timeframes for sampling because uncommon species or communities may be missed if a sufficient number of samples cannot be collected. Stratified random sampling ensures that all known types of areas will be sampled, but may still present challenges when some areas are difficult or impossible to access. Although subjective sampling can have an effect on analysis outcomes, it may still be an acceptable method for some otherwise untenable situations where specific hypothesis testing is not required.

2. This paper describes a method used to define strata to guide a single-season vegetation sampling effort over a large landscape. The Iterative Self-Organizing Data Analysis Technique (ISODATA) clustering algorithm was used to stratify the landscape by using several continuous data layers of abiotic attributes in a geographic information system (GIS). These strata were used in conjunction with a bedrock geology map and orthophoto imagery to choose potential field-sampling locations. Actual field-sampling locations were then chosen subjectively based upon the data needs of the project and accessibility.

3. This method provided an efficient means to guide the acquisition of vegetation information within major groups of abiotic strata over a large area during one field

[1]Coauthored by Kathryn Peterson and R. Douglas Ramsey

season. It also provided a method to determine where previously collected samples fell within the spectrum of abiotic attribute strata, which reduced the number of samples that needed to be acquired.

4. *Synthesis and applications.* The methods outlined in this paper provide a simple means to stratify a landscape based on abiotic attributes. By using these methods, sampling efforts can be allocated across a landscape efficiently. These methods are particularly useful where accessibility to some areas may be limited and/or where some areas may not be suitable for sampling. It also provides a means by which to determine where any previously collected samples lie within the range of abiotic attributes across an area. An additional benefit of the method is that it provides insight into the relationship between abiotic attributes and plant species distributions.

Introduction

Collecting field data over large geographic areas can present several challenges for ecologists. Completely random site selection is the best way to avoid sampling bias and, if enough samples are collected, provides a satisfactory estimate of sampling error (Jolly 1954). A basic assumption for standard statistical tests is that samples have been obtained randomly, giving each subset of the population an equal chance of being selected (Lájer 2007). When study areas are very large, many locations need to be sampled in order to adequately represent the variability of the landscape and provide an unbiased sample. If study areas are heterogeneous, a random sample that is not sufficiently large may undersample or completely miss uncommon types (Jensen 2005).

A stratified random sampling approach is often used to ensure that even uncommon areas will be sampled (Jensen 2005). Additionally, with this approach the

portion of natural variability due to differences between strata is automatically eliminated from the sampling error (Jolly 1954). Unfortunately over large study areas, stratified random sampling can still present challenges in areas that may be difficult or impossible to access. Also, for some studies, randomly chosen locations may not be appropriate sites for sampling. For example, if potential native plant species distributions are being studied, it would not make sense to collect data in recently seeded or otherwise modified locations.

Although subjectively chosen sampling locations have been proven to influence properties of experimental results (Hédl 2007), they still may be the best option for some studies where 1) survey areas are large and/or time/costs limit sampling intensity, 2) randomly selected sites might not provide data appropriate for the project, and 3) access to some areas is limited. As with all scientific inquiry, methods of sampling and data collection need to be driven by the objectives of the study.

This paper describes the approach taken to stratify and sample a large area with diverse terrain, varying accessibility, and varying suitability for sampling. Data collected via this process was utilized in the development of statistical models to predict potential plant species distributions. The accuracy of these models was assessed using contingency tables, therefore rigorous statistical tests comparing alternative hypotheses were not required.

Materials and methods

STUDY AREA

Rich County is located in the northeast corner of the state of Utah (USA) and is about 2811 km^2 (1085 mi^2) in size. Topography is quite varied, with western portions consisting of steep mountains having many narrow crests and valleys, while eastern portions contain broad basins, alluvial fans, piedmont plains, and pediment slopes from surrounding mountains. Overall, elevations range from about 1800 to 2800 m. The county falls within the rain shadow of the Bear River Range of the Wasatch Mountains; the highest portions of the county receive as much as 1300 mm of precipitation on average annually, while lower elevations receive as little as 260 mm. Mean annual air temperatures range from 2.3 °C to 5.8 °C; average monthly temperatures are fairly well correlated with elevation.

Most of the lower elevation and flatter portions of the county are in private ownership and managed for agricultural production and/or livestock grazing. Private lands make up about 56% of the county, the Bureau of Land Management (BLM) manages almost 25% of the land area, and the U.S. Forest Service manages just over 7%. The state of Utah manages just over 7% of the county as well, while the remaining 5% of the county is water – mostly the south half of Bear Lake.

The highest elevations of the county are forested, dominated by subalpine conifers and quaking aspen (*Populus tremuloides*). Most of the remainder of the county, except in riparian areas, is dominated by various sagebrush (*Artemisia* spp.) species. Some of the lowest elevation portions of the county have saline soils and are dominated by greasewood (*Sarcobatus vermiculatus*).

DATA NEEDS

The vegetation data collected from this sampling effort was used to produce potential plant species distribution maps for common species occurring in the county. Data needs required that as much of the county as possible be sampled in order to address the environmental variation across several species' distributions. The timeframe for this sampling effort was to be one summer field season, from mid May 2007 to late August 2007. Data were to be collected primarily by one person having experience sampling vegetation data in similar semi-arid environments.

GIS DATA LAYERS

Prior to applying the stratification procedure, variables to be used to stratify the landscape were identified. Since the focus of the project was to produce potential plant species distribution models, variables that related to soil water availability and temperature (both atmospheric and soil) were key. A conceptual model (Fig. 2.1) identifies variables that might be included in the species distribution modeling effort, and provides a guideline for choosing landscape stratification variables for the sampling effort.

Although some variables in the conceptual model are not easily acquired or computed, many are. Table 2.1 shows the data layers that were available or generated from available geospatial data. The digital elevation model (DEM), or *elev* dataset, was acquired from the U.S. Geological Survey's Seamless Data Distribution System (available at: http://seamless.usgs.gov). Once these data had been re-projected to match other data layers and clipped to a buffered county boundary, ArcMap (ESRI 2006) was

used to calculate slope (*slpd*) and slope curvature (*curv*). The specific catchment area (*sca*), or upslope contributing area, was computed using the TARDEM (Tarboton 2000) ArcMap plug-in which performs this calculation using the D-infinity method (Tarboton 1997). These values represent the size of the area that drains into each grid cell. The resulting *sca* values typically have an extremely wide range and large standard deviation. For this reason, the natural log of the specific catchment area was taken to produce the *ln_sca* variable (see Fig. 2.2). This is a common transformation for this variable; for example, in the computation of topographic wetness index (TWI) (Beven & Kirkby 1979) or terrain characterization index (TCI) (Park, McSweeney & Lowery, 2001) the log of the specific catchment area is used.

The climatic data layers (*ppt, tavg, tmax_sum, tmin_sum, tmax_win,* and *tmin_win*) were all acquired or computed from 1 km^2 grid Daymet datasets (available at: http://www.dayment.org) that had been scaled down to 90 m resolution (Zimmermann *et al.* 2007) using the procedures of Thornton, Running & White (1997). Although this downscaling does not make the data more accurate, it helps account for finer spatial variability over rugged terrain.

Monthly potential solar flux grids were generated using two Arc Macro Language (AML) programs developed by Zimmermann (shortwavc.aml and diffuse.aml, available at: http://www.wsl.ch/staff/niklaus.zimmermann/programs/aml.html#1) (12 June 2001) based on the work of Kumar, Skidmore, & Knowles (1997). These routines used the *elev* grid plus the latitude to compute potential clear-sky direct and diffuse solar radiation occurring on each pixel. The latitude of the town of Woodruff was used in the AMLs, as

this town marks the approximate north-south midpoint of the county. The direct and diffuse solar radiation grids were then summed for each month to produce 12 *sflux* grids.

The *brightn* variable was included because it appeared that neither the *geol* layer nor any of the other data layers were capturing apparent variations in vegetation and/or soil type that could be seen in orthophoto imagery. The *brighn* layer is the first feature (component) of a Tasseled Cap transformation (Crist & Cicone 1984) applied to a 3 October 2000 Landsat Thematic Mapper (TM) image. This feature highlights differences in soil characteristics such as particle size and distribution where soil is exposed; it was felt that this variability should be a component of the stratification. Inclusion of the second component of the Tasseled Cap transformation, "greenness", was considered as a stratification variable as well. However, when the greenness layer was inspected, it appeared to be highlighting agricultural areas, coniferous forest, and, to some degree, areas that had been seeded. Because the goal of stratification was to capture differences in abiotic attributes rather than differences in existing plant communities, this feature was not included.

A preliminary digital unpublished 1:100,000 scale polygon-based bedrock geology map for the county was acquired by special request from the Wyoming Geological Survey (Dover, 1995; Coogan & King, 2001; David W. Lucke, Wyoming State Geological Survey, personal communication). There were 59 different types of bedrock identified in the attribute table for the coverage (including water). These were grouped into 17 different types using primary bedrock material type names such as "sandstone", "limestone", or "conglomerate" that would be expected to have similar weathering characteristics and chemical properties (Table 2.2). The 17 groups were used

to produce the *geol* layer. Bedrock material types were sorted in descending order by the percentage of the county that each covered. The most common bedrock type, Wasatch Formation (grit/conglomerate/siltstone) composed about 62% of the county.

DATA PREPARATION

Landscape stratification consisted of a statistical clustering approach to objectively identify natural groupings of abiotic attributes. The Iterative Self-Organizing Data Analysis Technique (ISODATA) clustering algorithm is commonly used in remotely-sensed image analysis to stratify pixels into groups that have similar spectral characteristics across multiple bands (Jensen 2005). This clustering approach can also be used to stratify a landscape by partitioning a set of spatially-distributed abiotic attributes within a GIS environment. Automated clustering of landform attributes is not a new idea; the method was tried and compared with fuzzy classification for a 50 ha site in southwest Wisconsin (Irvin, Ventura & Slater, 1997). In that study, both methods showed promise for identifying landform elements, and the authors felt that the methods could be useful for statistical analyses and determination of sampling schemes. The approach was also utilized by Metzger *et al.* (2005) in their climatic stratification of Europe.

Prior to applying any clustering algorithm to these data, an analysis of elements containing several components (i.e. precipitation, temperature, and potential solar radiation grids) was performed to determine if the number of grids to be input into the clustering algorithm could be reduced. Reducing the number of data layers using principal components analysis (PCA) prior to applying a clustering algorithm was an approach used by Metzger *et al.* (2005).

A PCA was performed on the 12 monthly precipitation grids for Rich County; it showed that 99% of the variability of the grids could be accounted for in the first principal component (PC). Additionally, a correlation matrix of the original 12 precipitation grids showed that most of the grids were highly correlated with one another. From this analysis, it was determined that a simple average annual precipitation grid (*ppt_ann*) would be used rather than grid(s) produced from the PCA. A PCA of the 12 monthly *tavg* grids plus the *tmax_sum, tmin_sum, tmax_win,* and *tmin_win* grids indicated that over 99% of their variance could be accounted for with the first 3 principal components, so it was decided that the first 3 components (*temp_c1, temp_c2,* and *temp_c3*) would be used for landscape stratification.

When a PCA was performed on the 12 *sflux* grids it indicated that the first PC would capture more than 96% of their combined variability, and the first two PCs would account for more than 99% of their variability. An average daily solar flux grid (*sflux_avg*) was also computed by summing the 12 *sflux* grids and dividing by 365. Because the annual average solar flux grid appeared very similar to the first PC of the 12 *sflux* grids when displayed in the GIS software, the correlation between the two grids was computed. It was found that the two grids were correlated at 0.998 (r). For simplicity, it was decided that the annual average solar flux grid (*sflux_avg*) would be used for stratification. This grid appeared to be separating hot, dry southwest-facing slopes from cool, moist northeast-facing slopes as was expected. It was decided that the elevation grid would not be input into the clustering algorithm as previous analysis showed it to be highly correlated with temperature and somewhat correlated with precipitation grids.

One last alteration was made to the *curv* grid before proceeding further. It was noted that curvature values had a very wide range and non-normal distribution (Fig. 3), but that few grid cells (less than 2.26%) had values that deviated more than three standard deviations from the mean. It was felt that such a wide range might cause the clustering algorithm to create clusters that had curvature ranges that were far from "typical" or would require that more clusters to be created to account for the large range of the curvature variable. Because landforms with extreme convexity or concavity often have large changes in vegetation composition over small distances, these areas would not be appropriate for sampling. Sampling locations would be required to have relatively homogeneous vegetation cover over an area of at least 2825 m^2 (a circle with a radius of 30 m). For these reasons, the *curv* grid was adjusted so that values that were more than three standard deviations from the mean were assigned a value that was either three standard deviations above or below the mean as needed. The *geol* layer was not included in ISODATA algorithm as this was a categorical variable; bedrock material type was used for stratification after continuous variables had been clustered.

CLUSTERING

Before applying the clustering algorithm to these grids, the raw data values in each grid were converted to a standard deviation scale. This was done so that all grids would have equal weight when the ISODATA algorithm was applied. Next, a correlation matrix was computed to verify the statistical independence of the individual grids (Table 2.3). The correlation matrix indicated that none were highly correlated; the highest correlation (-0.84) was between the *temp_c1* and *ppt_ann*. The clustering algorithm was applied to the nine grids shown in Table 2.4.

The number of clusters to produce was an important consideration. The application of the ISODATA algorithm in a GIS produces a signature file containing means for each data layer (variable) in each cluster and a covariance matrix of variables within each cluster. Variances for each variable computed from standard deviations (which are in turn computed from covariance matrices) can be used to help decide how many clusters might be appropriate to produce to account for the variability across the datasets. In this approach, variances are summed across all variables in a cluster to obtain a total sum of variance for each cluster. These sums are again summed to obtain a total sum of variance. The total sum of variance is divided by the number of clusters to calculate the average sum of variance for that set of clusters (Table 2.5).

Typically, as the number of clusters increases, the average sum of variance decreases. The idea is to determine the minimum number of clusters acceptable to account for as much variability as possible. Though there is no objective criterion to determine the appropriate number of clusters to generate, a graphical representation of number of clusters vs. average sum of variance can be used to guide decision-making (see Fig. 2.4). The point at which the graph tends to "flatten out" indicates the optimal number of clusters that will balance the average variability within individual clusters with the number of clusters. This is similar to the way a scree plot might be used to help determine the appropriate number of PCs to keep from a PCA analysis. Based on the average sum of variance plot and consideration of the timeframe allowed for field-sampling, it was decided that 25 clusters would adequately stratify the landscape variability across Rich County.

Fig. 5 shows the 25-cluster map of Rich County produced from the ISODATA algorithm. Cluster numbers were sorted in descending order by total cluster area and colored systematically for ease of use. The map also shows the distribution of public and private lands within the county, as well as areas known to have been seeded based on a "treatments" data layer (Edwards, T.C., U.S. Geological Survey, Utah State University, personal communication). Actual field-sampling locations are also indicated on this map.

FIELD SAMPLING

The cluster map was further stratified with the simplified bedrock material map (*geol*) by creating a matrix between the 25 cluster groups and the 16 (non-water) bedrock material groups. Sampling locations were chosen opportunistically where the map indicated relatively large patches of cluster/bedrock type groups. Locations were usually located between 40 m and 250 m from roadways; data were not collected less than 30 m from roads as it was felt that the presence of roads could affect plant species composition or dominance patterns. An example of how the cluster map was used to guide sampling is shown in Fig. 2.6. If a particular location was determined to have a relatively homogeneous vegetation composition over a roughly circular area with a 30 m radius (approximately 2825 m^2) and was found to be dominated by native species, it was usually deemed adequate for sampling. Common plant species within a 30 m radius of chosen locations were recorded along with ocular estimates of foliar cover. Species were generally considered common if their foliar cover was >=1%. Taller-statured cool-season bunchgrasses that typically decrease on these rangelands with livestock grazing such as bluebunch wheatgrass (*Pseudoroegneria spicata*), needle-and-thread (*Hesperostipa comata* ssp. *comata*), and Letterman's needlegrass (*Achnatherum*

lettermanii), were considered common even if they had less than 1% canopy cover, especially if they could only be found under shrub canopies. Many of these rangelands have been heavily grazed in the past; historic grazing combined with recent dry years can cause these grasses to be reduced in abundance and vigor (Stoddart 1940). At most of the sampling sites there were fewer than four species within each life form (forbs, grasses, shrubs, or trees) that composed more than 1% of the foliar canopy cover. This project did not require careful measurement of cover or abundance; just noting the common species that were present and approximate cover was suitable for subsequent analysis.

After some initial sampling to determine the average time required for each sample and to estimate the total number of locations that could be sampled during the entire field season (we estimated between 200 and 225 samples could be collected), we calculated a target number of samples to obtain within each cluster/bedrock type group. Target numbers were proportional to the spatial area occupied by those groups within the study area (excluding water); this was done to reduce bias due to disproportionate sampling intensity (Cooper, McCann & Bunce, 2006). Because some cluster/bedrock type groups covered relatively small areas of the county, there would not be sufficient time to sample them. Where it seemed appropriate to further lump bedrock material types together to reduce the total number of strata, this was done. Table 6 shows the target number of samples for each combination of cluster and bedrock type(s); the total number of target samples was 216. As can be seen in Table 6, many cluster and bedrock type combinations were not targeted for sampling, and several combinations would have few samples. Though this was not an ideal situation, it was necessary to eliminate uncommon strata in order to be sure that more common types would be adequately

sampled. Initial field sampling also indicated that some cluster/bedrock type combinations would not be sampled proportionately because these areas were almost exclusively in private ownership and irrigated or altered.

Usually prior to each field excursion, several larger cluster/bedrock type patches that appeared relatively uniform on digital orthophoto imagery were located and UTM coordinates recorded (Fig. 2.6). Locations were then systematically visited over the course of the day. An effort was made to visit different parts of the county over the course of the summer, mostly on public land. Fortunately, some of the largest landholders in the county granted permission to collect data on their property, which helped ensure a more uniform distribution of sample locations. By the end of the 2007 field season data had been collected at 245 sites.

Results

The landscape stratification procedure provided an objective method to subdivide and efficiently sample a large landscape based on abiotic attribute groups. The method also provided a means to determine where pre-existing data could be utilized by indicating where pre-existing samples occurred in the spectrum of abiotic attribute strata. In the case of this project, clusters 22, 24, 25, and most of cluster 16 did not need to be sampled in 2007 because they had been sampled in 2001 using similar protocols for the Southwest Regional GAP Analysis Project (SWReGAP, Lowry *et al*. 2007).

To verify that the landscape had been sampled across the full range of abiotic attributes, the GPS coordinates of all of the samples collected in 2007 plus the additional 25 samples from 2001, a total of 270 sample locations, were used to create a point coverage in a GIS. This coverage was used to extract the values of all of the data layers

input into the ISODATA clustering algorithm at sampling locations. By comparing the sample distribution for each of the abiotic attributes with those of the attribute statistics across the entire county (Fig. 2.7), we could determine whether we adequately sampled the abiotic diversity across the county and whether we sampled abiotic attributes proportionally to their occurrence on the landscape.

Discussion

Based on the results shown in Fig. 2.7, the distribution of abiotic attributes was sampled similarly to the distribution of those variables across the landscape. Most of the sample density distribution curves match very closely with the distribution curves for grid values across the entire county. The three temperature PCs are the least similar. This may be because most of the highest-temperature areas are located in the lower parts of the county and are in private ownership and/or have been converted to agricultural fields. These areas were not sampled proportionately to their area within the county.

Tables 2.7 and 2.8 summarize the proportion of clusters and bedrock material types within the county respectively, and the number and proportion of samples collected within those groups. A few clusters were sampled disproportionately highly; these included clusters 16, 22, 24, and 25. As noted earlier, most of these samples were collected in 2001 for SWReGAP. Clusters 3 and 15 were undersampled because they occurred in small drainages which were difficult to sample because there were few places wide enough to have homogeneous vegetation cover over a 2825 m^2 area. Bedrock material types that were undersampled included High Level Alluvium and Gravel, which mostly occurred in areas of private ownership that have been altered for agriculture and/or livestock production. Another undersampled bedrock type was Dolomite &

Dolomite/Limestone. Undersampling of these types was due to their occurrence on steep slopes and in otherwise difficult-to-access areas.

This method of landscape stratification allowed for efficient canvassing of the entire county while taking into account several abiotic attributes that drive vegetation distribution. A great advantage to this method was that it allowed public lands to be sampled as much as possible, reducing time costs related to obtaining landowner permission to collect data on private lands, while still ensuring that data were sampled from sites that included the full spectrum of abiotic attribute groups. Application of the method was also an excellent way to obtain insights into plant species distribution patterns throughout the county in relation to abiotic factors. As sampling proceeded during the summer, the data collector became more able to predict what plant species might occur in particular cluster/bedrock types. The ability of this method to identify where SWReGAP samples could be utilized was a great benefit. For this project, 25 samples were obtained in this manner; collecting this number of samples in 2007 would have taken several days.

Although we considered the sampling effort based on this stratification technique successful, some improvements or changes could be made. One change that we could have made was the way in which bedrock material types were grouped. The original bedrock geology coverage had been separated into 17 different types by grouping based on primary bedrock material types. For many of these groups, there was no apparent soil or vegetation difference in the field. After reviewing the geology after the 2007 field season, we felt that it may have been sufficient to separate bedrock material types into only seven groups based primarily on their period of deposition. Groups of this type

would have included Eocene-Pliocene, Quaternary, Pre-Triassic, Triassic-Jurassic, Green River Formation, Cretaceous, and Water. We believe that these groups would have captured most of the differences in bedrock material types. This would have greatly simplified stratification, but would not have changed the sampling outcome.

A major concern in the field was the fact that the clusters did not differentiate some areas that were obviously quite different such as *Artemisia tridentata* (big sagebrush) and *Artemisia arbuscula* (low sagebrush) communities or other different types of communities that had similar abiotic features. Possibly the inclusion of more remotely sensed imagery layers, such as the inclusion of the "greenness" component of the Tasseled Cap Transformation, into the ISODATA algorithm might have separated these types into different clusters [though it has been noted that these species have similar spectral characteristics (Jakubauskas, Kindscher & Debinski, 2001)].

The fact that some areas with dissimilar vegetation were not partitioned into different clusters may have been due simply to the fact that an insufficient number of clusters were created using the ISODATA algorithm. We may have been inclined to create more clusters if there had been fewer categories of bedrock material types. Another issue could have simply been the resolution of the DEM and DEM-derived spatial data layers, particularly in areas with little topographic relief. We did consider using a 10 m DEM, but some processes are computationally intensive and the area of the county is approximately 2811 km^2.

It would be interesting to see the application of this stratification procedure to guide a more complete sampling effort, either with more personnel or over a longer period of time. With enough samples within each strata, more rigorous statistical

analysis such as hypothesis testing might be performed. Plant species distribution models developed using the data collected during this effort show promise based on accuracy assessments. Though the authors wish that there had been time to collect more samples, they can at least feel confident that they covered as much of the range of variability within the county as was possible given the limited timeframe allowed for sampling.

References

Beven, K.J. & Kirkby, M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Coogan, J.C. & King, J.K. (2001) Progress report: Geologic map of the Ogden 30' x 60' quadrangle, Utah and Wyoming, year 3 of 3. *Utah Geological Survey.*

Cooper A., McCann, T. & Bunce, R.G.H. (2006) The influence of sampling intensity on vegetation classification and the implications for environmental management. *Environmental Conservation,* **33**, 118-127.

Crist, E.P. & Cicone, R.C. (1984) A physically-based transformation of Thematic Mapper data – the TM tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing,* **GE22**, 256-265.

Dover, J.H. (1995) Geologic map of the Logan 30' x 60' quadrangle, Cache and Rich Counties, Utah, and Lincoln and Uinta Counties, Wyoming. *Utah Geological Survey.*

ESRI ArcMap 9.2 (2006) ESRI Inc. Redlands, CA.

Hédl, R. (2007) Is sampling subjectivity a distorting factor in surveys for vegetation diversity? *Folia Geobotanica*, **42**, 191-198.

Irvin, B.J., Ventura, S.J. & Slater, B.K. (1997) Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. *Geoderma,* **77**, 137-154.

Jakubauskas, M., Kindscher, K. & Debinski, D. (2001) Spectral and biophysical relationships of montaine sagebrush communities in multi-temporal SPOT XS data. *International Journal of Remote Sensing*, **22**, 1767-1778.

Jensen, J.R. (2005) *Introductory Digital Image Processing: A Remote Sensing Perspective*, 3rd edn. Pearson Prentice Hall, Upper Saddle River, NJ, 316 pp.

Jolly, G.M. (1954) Theory of sampling. *Methods of Surveying and Measuring Vegetation* (ed. D. Brown), pp. 8-18. Commonwealth Agricultural Bureaux, England.

Kumar, L., Skidmore, A.K. & Knowles, E. (1997) Modelling topographic variation in solar radiation in a GIS environment. *International Journal of Geographical Information Science*, **11**: 475-497.

Lájer, K. (2007) Statistical tests as inappropriate tools for data analysis performed on non-random samples of plant communities. *Folia Geobotanica*, **42**, 115-122.

Lowry, J.L. Jr., Ramsey, R.D., Boykin, K., Bradford, D., Comer, P., Falzarano, S., Kepner, W., Kirby, J., Langs, L., Prior-Magee, J., Manis, G., O'Brien, L., Pohs, K., Rieth, W., Sajwaj, T., Schrader, S., Thomas, K.A., Schrupp, D., Schulz, K., Thompson, B., Wallace, C., Velasquez, C., Waller, E., Wolk, B. (2007) Mapping meso-scale land cover over very large geographic areas within a collaborative framework: A case study of the Southwest Regional Gap Analysis Project (SWReGAP). *Remote Sensing of Environment*, **108**, 59-73.

Metzger, M.J., Bunce, R.G.H., Jongman, R.H.G., Mücher, C.A. & Watkins, J.W. (2005) A climatic stratification of the environment of Europe. *Global Ecology and Biogeography*, **14**, 549-563.

Park, S.J., McSweeney, K. & Lowery, B. (2001) Identification of the spatial distribution of soils using a process-based terrain characterization. *Geoderma*, **103**, 249-272.

Stoddart, L. A. (1940) Range Resources of Rich County, Utah. *Bulletin 291, Agricultural Experiment Station*. Utah State University, Logan, UT.

Tarboton, D.G. (2000) *TARDEM, A suite of programs for the Analysis of Digital Elevation Data*. Copyright (C) 2000, David Tarboton, Utah State University. Available at: http://www.engineering.usu.edu/cee/faculty/dtarb/tardem.html (Accessed 15 October 2007)

Tarboton, D.G. (1997) A New Method for the Determination of Flow Directions and Contributing Areas in Grid Digital Elevation Models. *Water Resources Research*, **33**, 309-319.

Thornton, P.E., Running, S.W. & White, M.A. (1997) Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, **190**, 214-51.

Zimmermann, N.E. (last updated 12 June 2001) *Tools for analyzing, summarizing, and mapping of biophysical variables*. Available at: http://www.wsl.ch/staff/niklaus.zimmermann/progs.html (accessed 15 October 2007).

Zimmermann, N.E. & Roberts, D.W. (2001) *Final Report of the MLP climate and biophysical mapping project*. Available at: http://www.wsl.ch/staff/niklaus.zimmermann/mlp/mlp_report.pdf (accessed 17 October 2007).

Table 2.1. Available or computed GIS data layers.

Variable	Description	Resolution
elev	digital elevation model	30 m
slpd	slope, in degrees	30 m
curv	curvature	30 m
ln_sca	natural log of specific catchment area	30 m
sflux (12)	monthly average potential solar flux	30 m
ppt (12)	monthly average precipitation	90 m
tavg (12)	monthly average temperatures	90 m
tmin_win	minimum average winter (January) temperature	90 m
tmax_win	maximum average winter (January) temperature	90 m
tmin_sum	minimum average summer (July) temperature	90 m
tmax_sum	maximum average summer (July) temperature	90 m
brightn	Tasseled Cap brightness	30 m
geol	bedrock material groups	1:100,000

Table 2.2. Bedrock material simplified groups and their constituent geologic map units. Continued on the following page.

Bedrock Material Groups/ Geology Map Units	Hectares
1 Grit/ Conglomerate/ Siltstone	172706.18
Wasatch Formation (middle and lower Eocene)	172706.18
2 High-level Alluvium	32662.54
Alluvium (Holocene and upper? Pleistocene), Flood plain deposits	28279.41
Quaternary and/or Tertiary high-level alluvium	461.28
Quaternary and/or Tertiary high-level alluvium/ Fowkes Formation (middle Eocene)	2302.69
Quaternary and/or Tertiary high-level alluvium?	619.41
Quaternary and/or Tertiary high-level alluvium?/ Fowkes Formation (middle Eocene)	797.34
Quaternary and/or Tertiary high-level alluvium?/ Fowkes Formation (middle Eocene)?	202.41
3 Streamside Deposits	21540.83
Alluvial and colluvial deposits	3524.85
Alluvial-fan deposits (Quaternary)	1798.44
Colluvium (Holocene)	85.01
Side-stream alluvium and fan deposits (Holocene and Pleistocene)	15418.97
Stream and fan alluvium	648.36
Stream-terrace deposits	65.19
4 Water	14515.71
Water	14515.71
5 Sandstone/ Siltstone	6696.75
Beirdneau Formation (Upper Devonian)	335.42
Fowkes Formation (middle Eocene)	6206.26
Fowkes Formation (middle Eocene)?	155.07
6 Limestone	5370.50
Blacksmith, Bancroft and Ute Limestones (Middle Cambrian)	420.65
Garden City Formation (Middle and Lower Ordovician)	1488.02
Garden City Formation (Ordovician)	381.42
Lodgepole Limestone (Lower Mississippian)	878.07
Twin Creek Limestone (Middle Jurassic)	1784.54
Twin Creek Limestone (Middle Jurassic) Boundary Ridge Member	66.51
Twin Creek Limestone (Middle Jurassic) Leeds Creek Member	236.29
Twin Creek Limestone (Middle Jurassic) Sliderock Member	114.99
7 Quartzite	5324.34
Brigham Quartzite (Middle and Lower Cambrian and Precambrian)	2843.29
Lower member of Geertsen Canyon Quartzite (Middle and Lower Cambrian and possibly upper Proterozoic)	624.42
Swan Peak Quartzite (Middle Ordovician)	521.76
Upper member of Geertsen Canyon Quartzite (Middle and Lower Cambrian and possibly upper Proterozoic)	1334.86

Table 2.2. Continued from the previous page.

Bedrock Material Groups/ Geology Map Units	Hectares
8 Gravel	5010.21
Gravel (Holocene and Pleistocene)	5000.02
Terrace gravel (Holocene and/or Pleistocene and/or Pliocene)	10.20
9 Dolomite	4673.73
Bighorn Dolomite (Upper and Middle Ordovician)	0.65
Brazer Dolomite (Upper and Lower Mississippian)	1723.06
Fish Haven Dolomite (Lower Silurian and Upper Ordovician)	156.64
Jefferson Dolomite (Upper Devonian)	788.92
Laketown and Fish Haven Dolomites	153.81
Laketown Dolomite (Silurian)	1804.43
Nounan Dolomite (Upper and Middle Cambrian)	46.22
10 Limestone/ Shale/ Siltstone	3015.56
Bloomington Formation (Middle Cambrian)	1467.71
Dinwoody Formation (Lower Triassic)	249.27
Green River Formation (lower Eocene)	536.43
Green River Formation (lower Eocene)?	762.15
11 Sandstone	2874.79
Nugget Formation (Lower Jurassic)	236.64
Nugget Sandstone (Jurassic? and Triassic?)	2638.15
12 Dolomite/ Limestone	1516.71
St. Charles Formation (Lower Ordovician and Upper Cambrian)	1374.20
St. Charles Formation (Ordovician and Upper Cambrian)	142.51
13 Quartzite/ Quartz Sandstone	1444.23
Wells Formation (Lower Permian and Upper and Middle Pennsylvanian)	1444.23
14 Other Fine	1342.37
Phosphoria Formation (Lower Permian)	662.27
Sage Junction Formation (Lower Cretaceous)	678.88
Thomas Fork Formation (Lower Cretaceous)	1.22
15 Other Coarse	889.99
Diamictite	65.57
Diamicton	145.95
Hams Fork Conglomerate Member (Upper Cretaceous)	234.99
Landslide and slump deposits	278.40
Saly Lake Formation (Pliocene and Miocene)	165.08
16 Moraine	796.60
Moraine (Pleistocene)	796.60
17 Dune deposits	643.80
Dune deposits (Holocene or Pleistocene)	443.18
Dune Sand and Loess	200.62

Table 2.3. Correlation matrix for the nine variables input into the ISODATA algorithm.

Layer	slpd							
curv	0.05422	curv						
ln_sca	-0.15684	-0.48185	ln_sca					
ppt_ann	0.4576	0.0289	-0.07083	ppt_ann				
temp_c1	-0.44834	-0.0397	0.13854	-0.84289	temp_c1			
temp_c2	-0.06112	-0.00061	0.05247	-0.26095	-0.00029	temp_c2		
temp_c3	-0.19307	-0.00086	-0.06132	-0.39272	-0.00122	0.00114	temp_c3	
sflux_avg	-0.24967	0.01893	-0.00422	-0.0888	0.07159	0.01076	0.0611	sflux_avg
brightn	-0.0507	0.04984	-0.16818	-0.20894	0.02515	-0.0902	0.42575	0.46392

Table 2.4. Data layers input into the ISODATA algorithm.

Variable	Description	Resolution
slpd	slope, in degrees	30 m
curv	curvature (within 3 sd of mean)	30 m
ln_sca	natural log of specific catchment area	30 m
ppt_ann	average of 12 *ppt* grids	90 m
temp_c1	1st principal component of 16 temperature grids	90 m
temp_c2	2nd principal component of 16 temperature grids	90 m
temp_c3	3rd principal component of 16 temperature grids	90 m
sflux_avg	average of 12 *sflux* grids	30 m
brightn	Tasseled Cap brightness	30 m

Table 2.5. Example of average total sum of variance calculation (cl = cluster).

cl	ppt	temp_c1	temp_c2	temp_c3	slpd	curv	ln_uca	sflux	brightn	sum of variance
1	96.61	91.55	17.30	18.61	74.36	67.18	142.21	46.54	55.76	53900
2	72.40	65.21	20.64	16.95	62.12	68.74	110.91	55.91	55.14	37257
3	77.87	77.03	17.93	19.39	61.44	65.92	75.02	45.36	70.45	33465
4	59.04	51.01	15.92	18.86	20.95	26.65	52.72	0.00	153.09	34061
5	40.28	44.59	23.52	14.77	43.47	42.31	38.38	50.72	63.78	16175
6	43.41	53.64	18.68	19.81	36.08	40.00	81.58	117.95	146.95	50569
7	64.15	56.00	19.61	17.29	63.90	60.24	94.50	77.66	55.04	33636
8	53.27	66.13	19.58	19.81	78.52	86.13	72.17	58.06	80.66	36654
9	67.30	58.35	21.59	16.72	59.11	85.15	83.06	36.05	62.91	31580
10	68.85	59.62	22.50	16.53	90.76	90.97	64.74	25.49	67.72	35014
									Average sum of variance:	36231

Table 2.6. Target number of samples to acquire for each cluster/bedrock type combination. As shown, several combinations did not make up a large enough percentage of the county to be targeted for sampling. Column totals indicate the percent of total area for each bedrock type group; row totals indicate percent of total area for each cluster. ** Column and row totals do not add to 100% due to exclusion of water from Bedrock material groups (group 4) and Clusters (cluster 9), and due to the fact that cluster areas, were initially computed up to 1 km beyond the county boundary, and then truncated to the county boundary.

	Bedrock material group									
Cluster	1	2	3	5, 11	6, 10	7, 13	8	9, 12	14-17	
1	9	17	4	3	0	0	2	0	0	14.18%
2	26	0	1	0	0	0	0	0	0	11.24%
3	7	8	4	1	0	0	1	0	0	7.94%
4	16	0	2	1	0	0	0	0	0	5.11%
5	10	0	1	0	0	1	0	0	0	4.74%
6	10	0	0	0	0	0	0	0	0	4.75%
7	11	0	0	0	0	0	0	0	0	4.83%
8	11	0	0	0	0	0	0	0	0	4.67%
10	6	0	1	1	1	0	0	0	0	3.41%
11	4	0	0	0	1	0	0	0	0	2.15%
12	6	0	0	0	0	0	0	0	0	2.91%
13	6	0	0	0	0	0	0	0	0	2.76%
14	2	3	2	0	0	0	0	0	0	2.91%
15	4	0	1	0	0	0	0	0	0	2.61%
16	3	0	0	0	1	0	0	1	0	1.87%
17	4	0	0	1	0	0	0	0	0	2.47%
18	3	0	0	0	1	0	0	0	0	1.96%
19	3	0	1	0	0	0	0	0	0	2.18%
20	3	0	0	0	0	1	0	0	0	1.88%
21	3	0	0	0	0	0	0	0	0	1.86%
22	2	0	0	0	0	0	0	0	0	1.53%
23	2	0	0	0	0	0	0	0	0	1.63%
24	2	0	0	0	0	0	0	0	0	1.41%
25	1	0	0	0	0	0	0	0	0	1.23%
	61.89%	11.46%	7.68%	3.43%	3.00%	2.43%	1.79%	1.79%	1.32%	**

Table 2.7. Proportion of clusters sampled. The 'Captured' column was calculated by dividing '% of Samples' by '% of County'. Cluster 9 is water.

Cluster	% of County	Samples	% of Samples	Captured
1	14.18%	35	12.96%	91%
2	11.24%	35	12.96%	115%
3	7.94%	12	4.44%	56%
4	7.75%	23	8.52%	110%
5	5.11%	15	5.56%	109%
6	4.74%	12	4.44%	94%
7	4.75%	14	5.19%	109%
8	4.83%	12	4.44%	92%
9	4.67%	0	0.00%	
10	3.41%	12	4.44%	130%
11	2.15%	7	2.59%	120%
12	2.91%	9	3.33%	115%
13	2.76%	7	2.59%	94%
14	2.91%	7	2.59%	89%
15	2.61%	5	1.85%	71%
16	1.87%	11	4.07%	217%
17	2.47%	10	3.70%	150%
18	1.96%	5	1.85%	95%
19	2.18%	5	1.85%	85%
20	1.88%	7	2.59%	138%
21	1.86%	4	1.48%	80%
22	1.53%	8	2.96%	193%
23	1.63%	4	1.48%	91%
24	1.41%	6	2.22%	157%
25	1.23%	5	1.85%	150%

Table 2.8. Proportion of bedrock material types sampled. The 'Captured' column was calculated by dividing '% of Samples' by '% of County'.

Bedrock Material Group		% of County	Samples	% of Samples	Captured
1	Grit/Conglomerate/Siltstone	61.89%	198	73.33%	118%
2	High-level Alluvium	11.46%	14	5.19%	45%
3	Streamside Deposits	7.68%	26	9.63%	125%
4	Water	5.20%	0	.	.
5, 11	Sandstone/Siltstone & Sandstone	3.43%	9	3.33%	97%
6, 10	Limestone & Limestone/Shale/Siltstone	3.00%	10	3.70%	123%
7, 13	Quartzite & Quartzite/Quartz Sandstone	2.43%	6	2.22%	92%
8	Gravel	1.79%	3	1.11%	62%
9, 12	Dolomite & Dolomite/Limestone	1.79%	1	0.37%	21%
14-17	Other Fine, Other Course Moraine, & Dune Deposits	1.32%	3	1.11%	84%

Figures

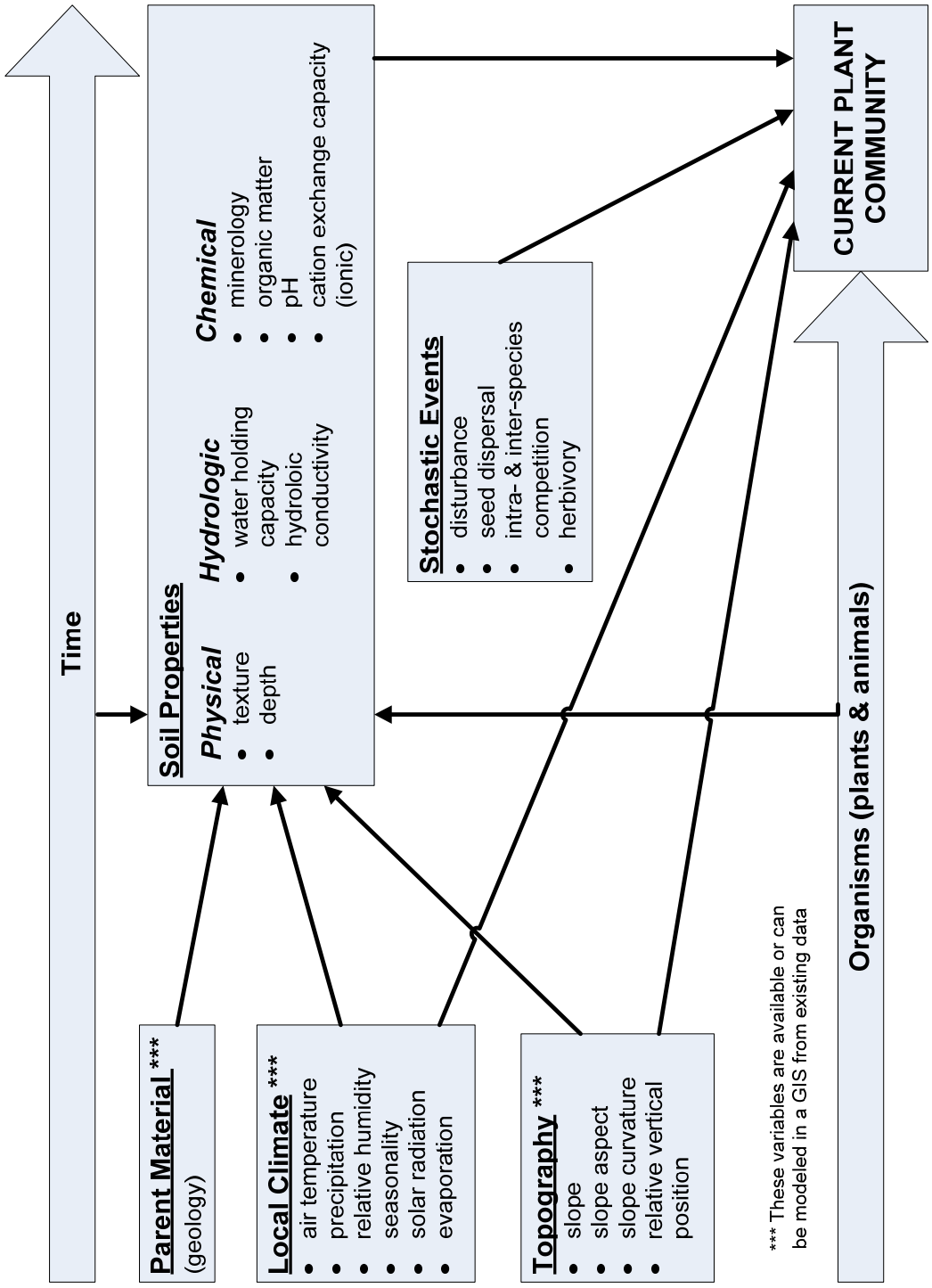

Fig 2.1. Conceptual model showing variables which drive plant species distributions in semi-arid environments.

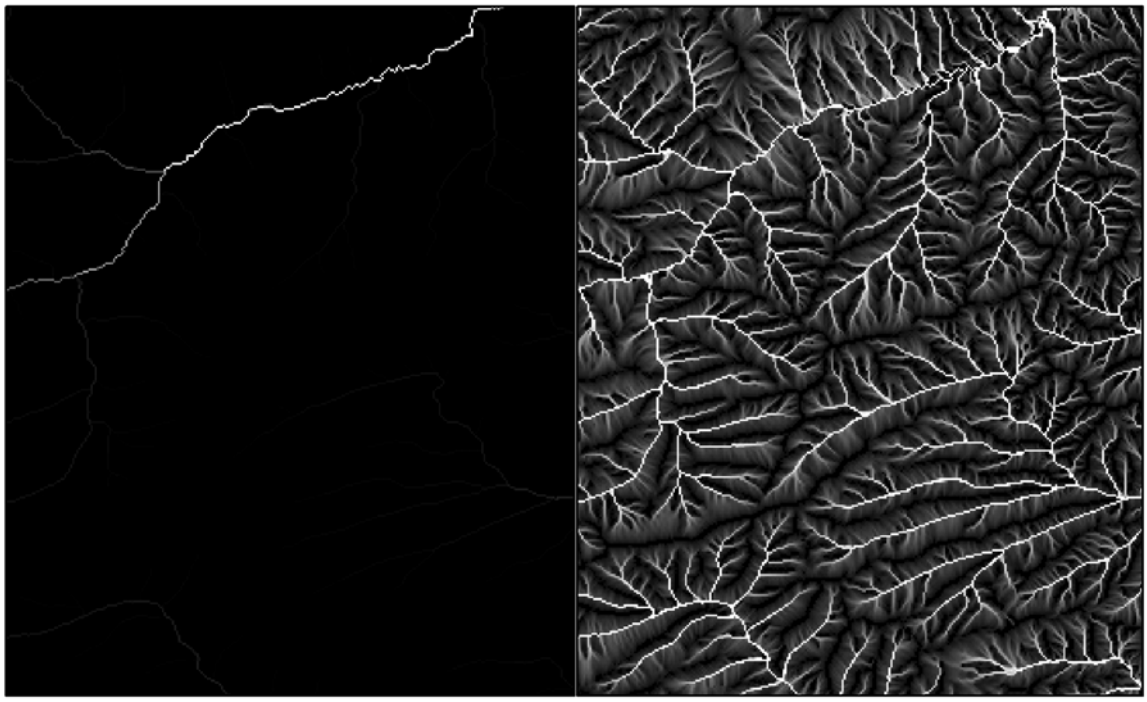

Fig. 2.2. Illustration of specific catchment (*sca*) raw values compared to natural logarithm transformed *sca* values (*ln_sca*). The raw *sca* grid on the left had a minimum of 30, maximum of 18,697,444, mean of 157,973, and standard deviation of 1,656,326. The *ln_sca* grid on the right had a minimum of 3.40, maximum of 16.74, mean of 5.27, and standard deviation of 2.

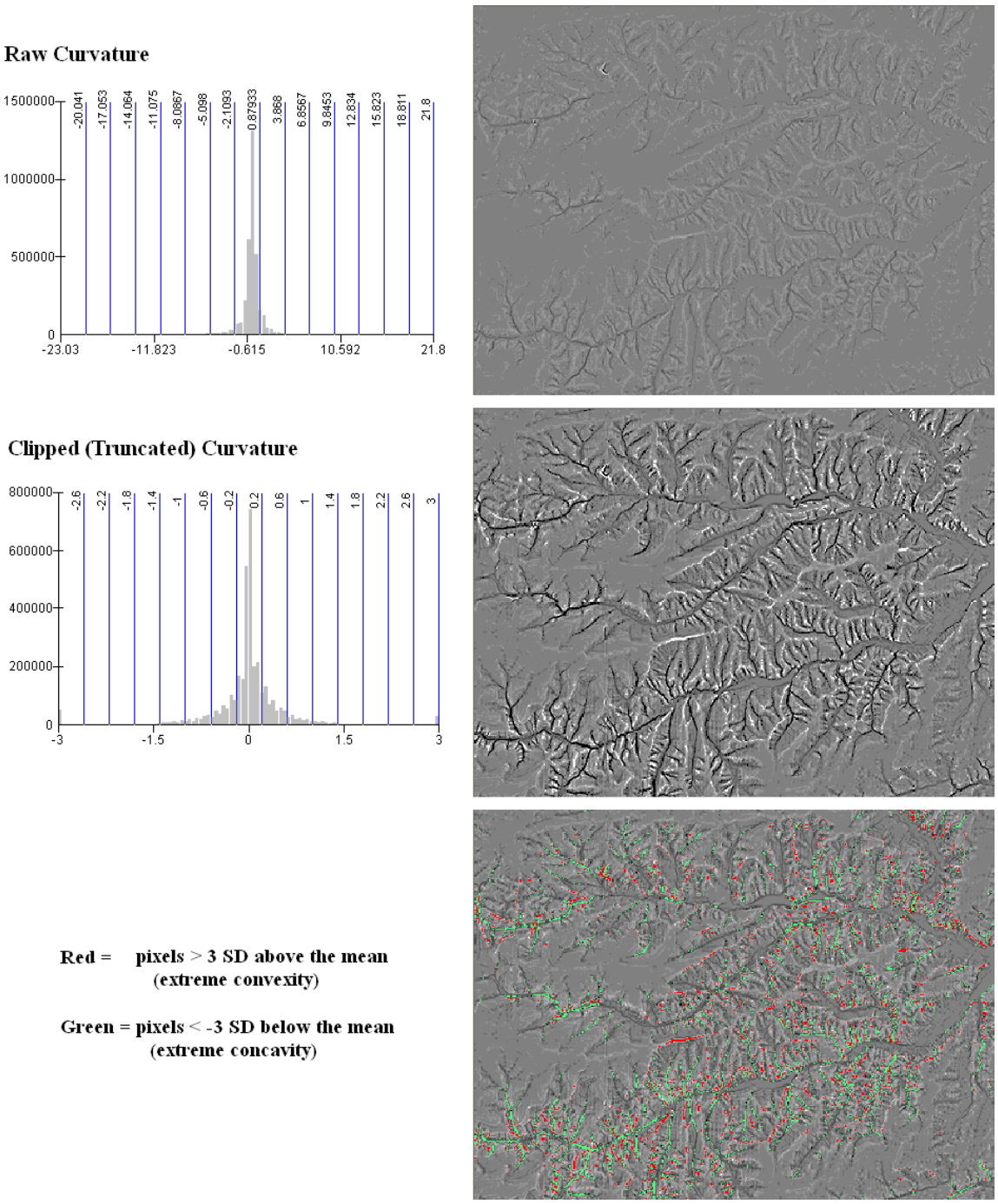

Fig. 2.3. Illustration of why slope curvature was cut off at three standard deviations above or below the mean. Raw curvature values ranged from -11.29 to 10.68, had a mean of 0, and a standard deviation of 0.48. Less than 2.27% of pixels fell above or below three standard deviations of the mean. Grayscale of maps is in 15 equal intervals.

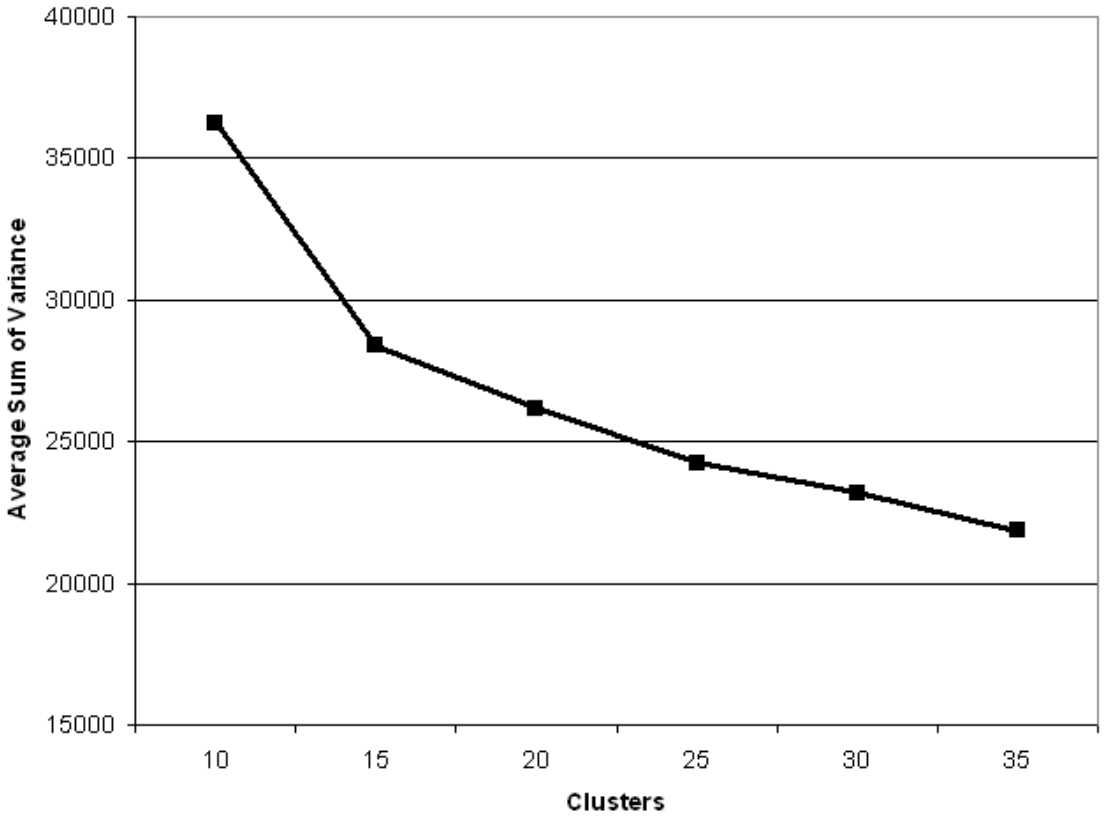

Fig. 2.4. Average sum of variance plot.

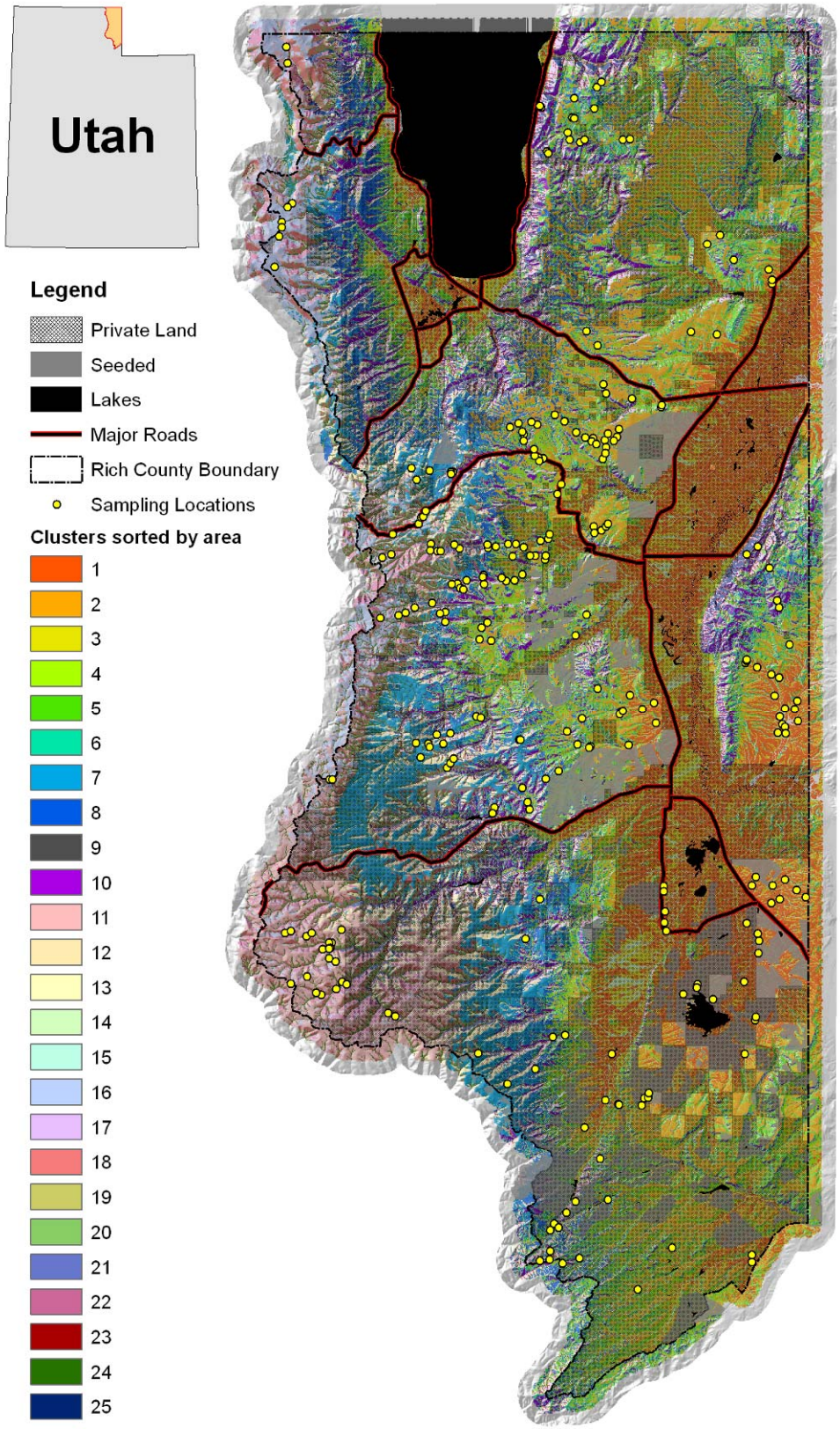

Fig. 2.5. Clusters and sampling locations in Rich County.

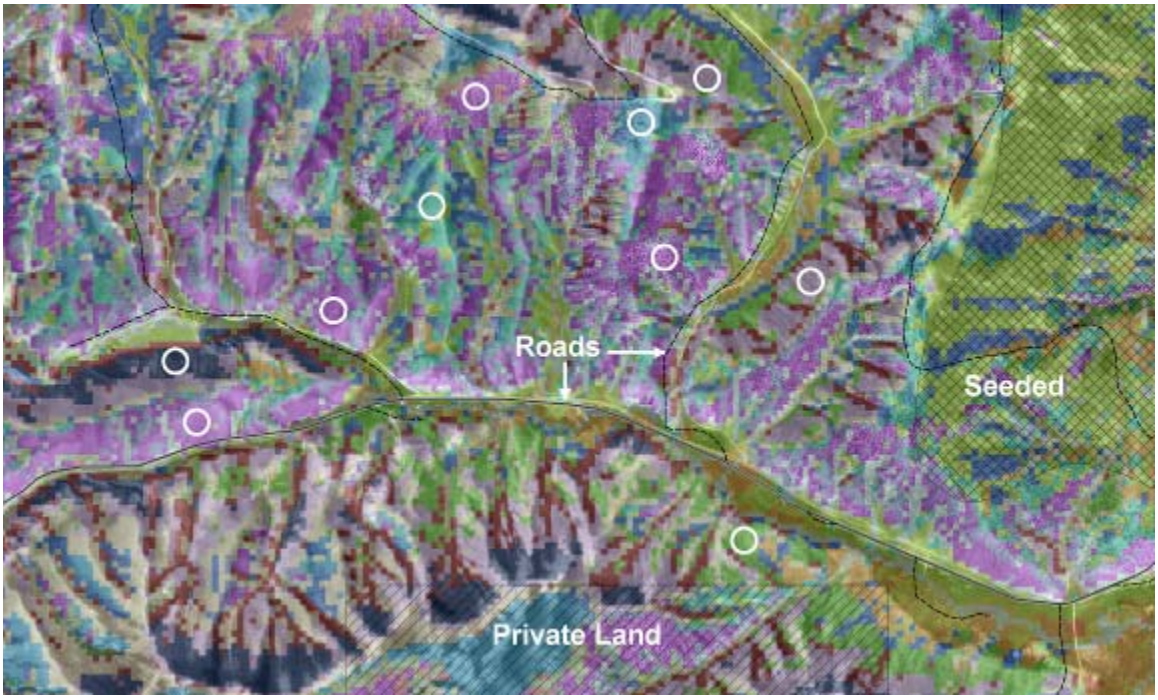

Fig. 2.6. Illustration of how the cluster map was used to help guide field sampling. In this example, cluster colors are draped over an orthophoto image; roads, land ownership and rangeland treatment coverages are also shown. Relatively large areas of clusters/bedrock types were targeted as potential sampling locations. Areas such as those indicated by white circles might be visited. Sites were chosen to be between 40 and 250 m from roads.

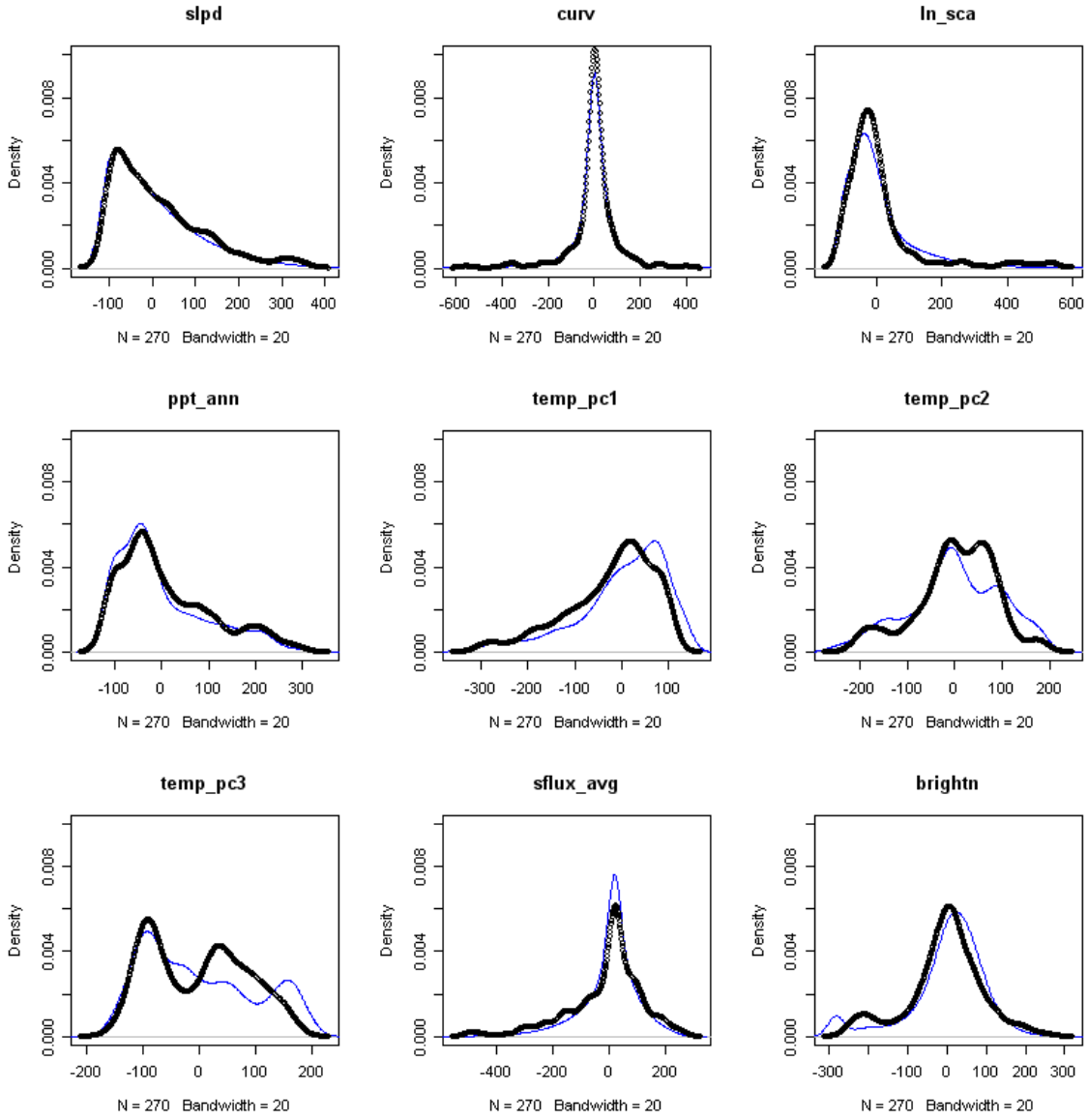

Fig 2.7. Density distribution of abiotic attribute values of sample data compared to the density distribution of abiotic attribute values across the county. Blue lines indicate the distribution of values across the county; black circles show the distribution of abiotic attributes at sample locations. Attribute values are in standard deviation scale × 100.

CHAPTER 3

MODELING THE POTENTIAL DISTRIBUTION OF COMMON PLANT SPECIES USING GIS-DERIVED ABIOTIC ATTRIBUTES AND LANDSAT TM IMAGERY IN RICH COUNTY, UTAH[1]

Abstract

Georeferenced field data were used to develop logistic regression models of common plant species distributions throughout Rich County, Utah (USA). Models were developed for 38 species or species groups which are common in various plant communities in the county. Predictor variables for these models included elevation, slope, slope curvature, specific catchment area, average annual precipitation, average monthly temperature, potential monthly solar flux, and a Landsat TM image. Principal components analysis (PCA) was done on monthly temperature grids, solar flux grids, and Landsat imagery to reduce the number of potential model variables. Model variables were selected for each species by a forwards/backwards stepwise procedure performed on 100 subsets of the training data; each subset consisted of 80% of the training data. Once model variables were selected, model coefficients and a "maximum sensitivity + specificity" (MS+S) threshold were computed using the entire training dataset. Model coefficients were applied to data layers in a geographic information system (GIS) to produce logit-scale output, which was converted to odds-scale. Odds estimates were normalized using the MS+S value so that the threshold between common and not-common classes was 0.5 for all species.

[1]Coauthored by Kathryn Peterson and R. Douglas Ramsey

An independent dataset derived from samples collected for two unrelated projects was used to evaluate 28 of the model outputs. Average estimated model sensitivities (predicting that species would be common in locations where they were common in the evaluation data) and overall correct classification rates were 0.626 and 0.683 respectively. Because of known issues with the independent dataset, accuracy estimates were also produced using an internal (bootstrap) cross-validation procedure. Using this method, species with more than 10 'common' occurrences had average estimated model sensitivities and overall correct classification rates of 0.734 and 0.813, respectively.

Although the independent dataset was not collected with this type of modeling effort in mind, we wanted to see if models could be improved by adding the independent data to the original training data. To this end, two additional models were created for each species. The first additional models used all of the original training data plus the independent evaluation dataset ('all of the data') to select model parameters and generate model coefficients. The second additional models used the same parameters as the original models, but generated parameter estimates using all of the data. Based on bootstrap cross-validation-estimated accuracy statistics, model parameters were best chosen using only the original training dataset, but parameter estimates were often better when computed using all of the data. The average sensitivity and correct classification rate estimates of the models with the best accuracy estimates were 0.795 and 0.840, respectively.

1. Introduction

Ecological types across diverse landscapes are most naturally differentiated by plant species compositions and landform attributes. Because soil characteristics and other abiotic attributes strongly affect plant community composition and distribution, the Natural Resources Conservation Service (NRCS) bases their Ecological Site Descriptions (ESDs) on abiotic attributes. Ecological sites are defined by the NRCS as "a distinctive kind of land with specific physical characteristics that differs from other kinds of land in its ability to produce a distinctive kind and amount of vegetation" (U.S. Department of Agriculture, National Soil Survey Handbook, Part 622.07, 2007).

Plant community species composition involves not only the identity and number of species, but abundance of each species as well. Species composition arises partly from deterministic processes linking habitat characteristics to species-specific niches, and stochastic processes such as seed dispersal (Ozinga et al., 2005). Plant communities are usually composed of a small set of relatively abundant (dominant or common) species mixed with a larger number of minor species (Hall, 1992; Walker et al., 1999).

Dominant species are considered to be the best adapted species for the local suite of abiotic and biotic factors, and serve to maintain ecosystem function (Walker et al., 1999). Vegetation associations are the result of seed availability and environmental selection, and environments are principally defined by climate and soil, altered by physiographic and biotic processes. Where climatic and physiographic changes are slow, continued propagule inputs and species interactions tend to produce relatively stable and static vegetation structures (Gleason, 1926). Predictive modeling of species distributions relies on the assumption of an equilibrium between biotic communities and abiotic

factors (Guisan and Theurillat, 2000). This premise is necessarily restricted to limited temporal scales, and can not be applied where communities are undergoing rapid succession or change.

Through the use of readily available spatial data layers and geo-referenced vegetation data for training, correlative (or empirical) statistical models describing the potential distributions of plant species can be developed. Potential species distribution maps generated from these models can be used to help determine the potential spatial distribution of plant communities across a landscape. This work was undertaken to help determine the fine-scale distribution of ecological site types across the county so that ESD correlations to soils in Rich County could be reviewed objectively using a data-driven process.

2. Materials and Methods

2.1. Study area

Rich County is about 2811 km^2 (1085 mi^2) in size and is located in the northeast corner of the state of Utah. Elevations range from about 1800 to 2800 m. The western portion of the county is bordered by the eastern side of the Bear River Range of the Wasatch Mountains, which cause a rain-shadow effect across the county. Lower elevations average as little as 260 mm of precipitation annually, while the highest elevations in the northwest part of the county average as much as 1300 mm. Mean annual air temperatures range from 2.3 °C to 5.9 °C.

The highest elevations of the county are forested; major species include *Pseudotsuga menziesii* (Mirbel) Franco [Douglas-fir], *Abies lasiocarpa* (Hook.) Nutt. [subalpine fir], *Pinus contorta* Dougl. ex Loud. [lodgepole pine], and *Populus*

tremuloides Michx. [quaking aspen]. At mid elevations, drier, rockier slopes have juniper, primarily *Juniperus osteosperma* (Torr.) Little [Utah juniper], while more moist slopes are dominated by the shrubs *Artemisia tridentata* Nutt. ssp. *vaseyana* (Rydb.) Beetle [mountain big sagebrush], *Amelanchier utahensis* Koehne [Utah serviceberry], *Symphoricarpos oreophilus* Gray [mountain snowberry], and occasionally *Purshia tridentata* (Pursh) DC. [antelope bitterbrush]. Lower mid-elevation slopes often have areas dominated by *Artemisia nova* A. Nels. [black sagebrush], or *Artemisia arbuscula* Nutt. ssp. *longiloba* (Osterhout) L. Shultz [early low sagebrush]. Moderate slopes above broad plains at lower elevations are mostly dominated by *Artemisia tridentata* Nutt. ssp. *wyomingensis* Beetle & Young [Wyoming big sagebrush], and occasionally include *Krascheninnikovia lanata* (Pursh) A.D.J. Meeuse & Smit [winterfat]. Broad, flat plains in the lowest portions of the county have patches that vary from *Artemisia tridentata* Nutt. ssp. *tridentata* [basin big sagebrush], to *Sarcobatus vermiculatus* (Hook.) Torr. [greasewood], to *Carex* L., *Juncus* L., and *Salix* L. [sedge, rush, and willow] communities.

2.2. Field sampling

During the summer of 2007 Rich County was stratified into 225 abiotic attribute strata. Abiotic attribute groups were based on annual average precipitation, the first three principal components of a PCA done on 12 average monthly temperature grids plus maximum and minimum summer and winter temperature grids (four additional temperature grids), degree of slope, slope curvature, the natural log of the specific catchment area, annual potential direct plus diffuse solar radiation, and the brightness component of a Tasseled Cap transformation (Crist and Cicone, 1984) of an early

October Landsat Thematic Mapper (TM) image of the area (Chapter 2, Using GIS-Derived Clusters Of Abiotic Factors to Guide a Limited Field-Sampling Effort). Abiotic attribute groups were further subdivided by bedrock geology types to generate the final sampling strata. Due to limited time for field sampling, only the abiotic strata covering the largest percentage of the county were sampled. Vegetation data were collected at 264 sites; an attempt was made to sample strata proportionally to the strata's total area over the entire county. Sampling locations were distributed throughout the county, mainly on public lands, where some diversity of native plant species existed. These data were augmented by samples collected at 25 subalpine conifer-dominated sites 2001 for the Southwest Regional GAP Analysis Project (SWReGAP; Lowry et al., 2007), and 65 low- to mid-elevation sites sampled in 2006 for an initial field-based attempt to correlate soils to vegetation, bringing the total number of sampling locations to 354. At each location, common species were recorded along with ocular estimates of their foliar cover and the GPS coordinate; species with a foliar cover >=1% were considered common. These data were input into a customized Microsoft Access™ database. An additional 12 non-rangeland (8 water and 4 bare ground) locations were added to the database by sampling from orthophoto imagery to bring the total number of training samples to 366. It was hoped that the addition of non-rangeland samples would be useful in the creation of logistic species distribution models by identifying suites of abiotic characteristics that did not support specific species. By using GPS coordinates from the training sample database, a point coverage was created so that GIS data layers could be sampled at each location.

2.3. Development of logistic models of potential common species distributions

2.3.1. Spatial data layers

Based on a conceptual model describing factors and processes that affect plant species distributions (Fig. 3.1), the following data were used to develop logistic models of potential common species distributions:

- A digital elevation model (DEM)
- PRISM (Parameter-Elevation Regressions on Independent Slopes Model, available at www.prismclimate.org) monthly average precipitation and temperature grids
- A Landsat TM image

The DEM was acquired for the Rich County area from the United States Geological Survey (USGS) Seamless Data Distribution System (http://seamless.usgs.gov) and clipped to a 1 km buffered county boundary. The DEM was used to compute degree of slope (*slpd*) and slope curvature (*curv*) grids. Specific catchment area (*sca*) was extracted from the DEM using a method developed by Tarboton (1997). The resulting *sca* values typically have an extremely wide range and large standard deviation (see Fig. 3.2). For this reason, the natural log of the specific catchment area was taken to produce the *ln_sca* variable. This is a common transformation for this variable; for example, in the computation of topographic wetness index (TWI) (Beven and Kirkby, 1979) or terrain characterization index (TCI) (Park et al., 2001) the log of the specific catchment area is used.

Monthly potential solar flux grids were generated using Arc Macro Language (AML) programs developed by N.E. Zimmermann (shortwavc.aml and diffuse.aml,

available at: http://www.wsl.ch/staff/niklaus.zimmermann/programs/aml.html#1) based on the work of Kumar et al. (1997). Hourly solar flux grids were calculated and integrated by month to produce 12 solar flux (*sflux*) grids. The *sflux* grids were subjected to a principal components analysis (PCA) to reduce the number of potential data layers. The first principal component (PC1) accounted for 96.19% of the variability of solar flux within the study area; PC2 accounted for 3.61% of the variability. Factor loadings for PC1 were almost equal for all months except the months in the early summer when solar angles are at their highest (see Table 3.1). The second component, in contrast, had its highest three loadings in the high solar angle months. It was decided that both PC1 and PC2 would be used for logistic modeling. The remaining components were not considered further.

A PCA was also performed on the 12 monthly PRISM average precipitation grids. The first principal component accounted for more than 99% of the variability of the grids over the study area. This principal component also had a correlation greater than 0.99 with total annual average precipitation, so it was decided that a simple annual average precipitation grid would be used for modeling. This grid (*ppt_ann*) was created by summing monthly average precipitation grids.

Twelve average monthly temperature grids were created from PRISM data by averaging downloaded monthly average daily minimum and maximum temperature grids. A PCA was done on the monthly average temperature grids and it was found that 97.75% of the variability of these grids over Rich County was accounted for by the first three principal components. Factor loadings beyond the first three components did not appear to have an interpretable pattern, and so were discarded (see Table 3.2).

Various other PRISM-derived grids were reviewed including annual minimum and maximum temperature and difference between annual average maximum and minimum temperatures. Additionally an aridity index was computed by taking average annual precipitation and dividing by potential evapotranspiration, which had been modeled from DEM and temperature data using a program developed by Zimmermann (etp_jen.aml, available at: http://www.wsl.ch/staff/niklaus.zimmermann/programs/aml.html#3) based on an empirical equation developed by Jensen and Haise (1963). A seasonality index was created by dividing the total average precipitation for November through January (the three wettest months) by total average precipitation for June through August (the three driest months). All of these additional grids were highly correlated (greater than 0.90 (r)) with elevation and/or other grids already to be used for modeling (see Table 3.3), so it was decided that elevation would be included in logistic modeling procedures as a surrogate.

The seven spectral bands, including the thermal band, from a cloud-free 3 October 2000 Landsat TM image covering Rich County was acquired from the Intermountain Region Digital Image Archive Center (IRDIAC, http://earth.gis.usu.edu). The bands were clipped to a 2 km buffered boundary of Rich County and a PCA performed. The analysis showed that the first three principal components accounted for 98.8% of the variability of all of the bands (see Table 3.4). Therefore PCs 1 – 3 were used for modeling, while PCs 4 - 7 were discarded.

The 13 data layers used in modeling were converted to a standard deviation (or Z-score) scale by subtracting the layer's mean from actual pixel values and dividing by the layer's standard deviation (Hamilton, 1991). This was done so that any interactions

between variables would not be overly influenced by variables with larger measurement units. Standardized values greater than four were set to a value of four, and those less than -4 were set to a value of -4; this ensured that all data layer values were within four standard deviations of the mean. This was done because a few variables, particularly slope (Fig. 3.3), slope curvature, and solar flux PCs, had several pixels with outlying values. These outlying values appeared to be negatively influencing accuracies of test models. None of the data layers had more than 1.2% of pixels with standard deviations greater than ± 4. A correlation matrix was calculated for the 13 grid datasets to determine statistical independence of each variable (see Table 3.5).

2.3.2. Selecting logistic regression model variables

Several methods to determine which variables to include in the species distribution models were evaluated. The first issue considered was which interaction variables might be important. The solar flux, average temperature, and Landsat PC variables were considered for use as interaction variables as well, but only solar flux and average temperature first principal components were used. The highest values for the first PC of the solar flux grids indicted the sunniest, driest locations, generally facing southwest; lowest values indicated the opposite. Similarly, the high values for the first PC of the average temperature grids indicated higher annual average temperature; this PC was correlated with the annual average temperature grid computed from PRISM data at 0.865 (r).

It was decided that elevation would not be used as an interaction variable because it was highly correlated (0.899) with annual average precipitation, and precipitation would make more sense biologically than elevation. All possible combinations of the

following variables were considered for use in logistic regression models: *slpd*, *curv*, *ln_sca*, *ppt*, *tavg_pc1*, and *sflux_pc1*. A correlation matrix computed for all 15 of the resulting interaction variables (Table 3.6) showed that none were correlated at greater than -0.761 (r).

2.3.3. Building logistic regression models

The 13 abiotic data layers, including the three Landsat PC variables, were sampled in a GIS using the 366 sampling locations in the training database. Training data were associated with selected species which were coded 1 if common, 0 if not, to provide 366 training data samples for each species. Generally species with a foliar cover >= 1% were considered common. Known decreaser grasses were considered common if they were simply present regardless of percent cover; it was assumed that these grasses had been reduced to extremely low cover by historic or recent livestock grazing. Training data were imported into the R statistical software (freeware available at http://www.r-project.org) for analysis. For each species, the following procedure was done in R:

1) Eighty percent of the data were randomly selected from the training data.
2) Logistic regression parameters were fit to the randomly-selected training data using all of the model variables and interactions shown in Table 7.
3) A forwards/backwards stepwise regression procedure was used to reduce the number of model variables. The resulting step-selected model was stored in a table. The names of the variables used in the model were also stored as a list of separate data elements.

4) Steps 1-3 were repeated 100 times for each species. Following each iteration, the resulting step-selected model was added to the table, and the names of the variables used in that model were appended to the list in step 3.
5) After 100 iterations, the total number of times each variable was selected by the stepwise procedure was calculated, and the stepwise-selected models were sorted, grouped, and counted to see which had been produced most often.

Output from the R software was reviewed to see which of the model variables had been chosen by the stepwise process most frequently. Variables that had been selected more than 50% of the time were considered important, i.e. they should be used in the final model for the species. The table of step-determined models was also reviewed. If a model had been selected more than 10 times (out of 100), it was seriously considered as a potential final model. We felt it was important to consider not only the most commonly-selected variables, but also consider commonly selected suites of model variables. This is because suites of variables are "proven entities" – they were selected as a group to model the species – having potentially redundant variables eliminated and/or including variables that *together* fit the data best. On the other hand, no judgment could be made as to how individual variables would work together. Based on the variables chosen most often and the most common step-determined models, a final model for the species was determined. For almost every species the final model was one selected from the summary of step-determined models, usually one that had been determined relatively frequently.

Finally all training data (366 records) were used to determine parameter estimates for the final model predictor variables for each species. A "maximum sensitivity +

specificity" (MS+S) threshold was calculated using the "optimal.thresholds" function in R's PresenceAbsence package. A confusion matrix was then produced using the training data classes (common or not-common) and model predictions to provide an initial assessment of model accuracies.

The MS+S threshold determined for each species is where (sensitivity + specificity) / 2 (S+S/2) is determined to be greatest. For the purposes of this project, sensitivity is the percentage of the time that the species was predicted to be common when it was common in the training data. Specificity is the percentage of time the species was predicted to be not-common when it was not-common in the training data. The use of this threshold rather than simply using 0.50 was done to increase model accuracies. Logistic regression in particular produces estimates biased towards the larger group (Fielding and Bell, 1997); in the case of our training data, the larger group was almost invariably the not-common class, so modeled probabilities were biased towards that class. Various methods based on confusion matrices can be used to determine appropriate threshold values for the training data and the intended use of the model. For this application we wanted to be sure that models were identifying where species were common as accurately as possible (had high sensitivity), but still maintain high model specificity. Because training data usually contained many more occurrences of not-common than common, simply maximizing the correct classification rate would not achieve this goal.

The final logit probability models produced raster layers that were converted to an odds scale by computing $odds = 1 / (1 + e^{-logit})$. We developed two formulae that were used to adjust the odds estimate outputs to normalize the threshold between common and

not-common to 0.5 for all species, while still maintaining odds estimates between 0 and 1. Odds estimate values below the MS+S threshold value were adjusted using the formula:

$$Adjusted\ Odds = odds \times (0.5 / MS{+}S)$$

Odds estimate values above the MS+S threshold were adjusted to an odds value between 0.5 and 1 by using the formula:

$$Adjusted\ Odds = odds \times (0.5 / (1 - MS{+}S)) + (1 - (0.5 / 1 - MS{+}S))$$

For most of our model outputs, the first formula served to increase probability values that were below the MS+S threshold up to 0.5, while the second served to compress values between the MS+S threshold and 1 so that they would range between 0.5 and 1 (see Fig. 4). Standardizing thresholds was necessary for future work involving the analysis of multiple species models as a unit. The map outputs from six species (*Artemisia tridentata* ssp. *wyomingensis* [Wyoming big sagebrush], *A. tridentata* ssp. *vaseyana* [mountain big sagebrush], *Sarcobatus vermiculatus* [greasewood], *Pseudoroegneria spicata* [bluebunch wheatgrass], *Achnatherum lettermanii* [Letterman's needlegrass], and *Achnatherum hymenoides* [Indian ricegrass] are shown in Fig. 3.5.

Models were developed for 16 shrub species, two tree species, one tree group (subalpine conifers), 18 grass or grass-like species, and one grass/grass-like group (wetland sedge/rush/grass) for a total of 38 models. These species (and groups) are shown in Table 3.8. As can be seen in the table, several species had few occurrences of commonness. Models were developed for these despite their infrequent commonness because they were felt to be important for delineating ecological types in Rich County.

The frequency with which modeling variables were selected for final species models and their average P-values in fitted models are shown in Table 3.9.

It should be noted here that *A. tridentata* ssp. *tridentata* (basin big sagebrush) was divided into two groups for modeling, as there appeared to be two types of locations in which this species occurred. As expected, this species occurred in low-lying areas with added run-on moisture, but it also occurred in the south end of the county in upland locations with sandy-surfaced soils. Before validation, the two groups of basin big sagebrush were combined into one map by finding the maximum probability value of the two separate model outputs for each pixel. Also it should be noted that only sites with old-growth *Juniperus osteosperma* (Utah juniper) were used for modeling; sites with strictly younger (pointy-topped) junipers were not used. This was done because the NRCS considers younger juniper growing on deep soils or soils without many rock fragments to be invasive. Utah juniper models were not validated because independent validation data did not specify whether junipers were old-growth or young.

2.4. Model evaluation

2.4.1. Model evaluation using an independent dataset

Model evaluation data incorporated data from two independent sources [I here use the term *evaluation* rather than *validation* following Guisan and Zimmermann (2000)]. One source was the Bureau of Land Management (BLM), which had collected data for Ecological Site Inventory during the summers of 2005 and 2006 in the north-central part of the county. These data consisted of 67 samples where annual production (by dry weight) of species was estimated by either NRCS clipping-and-weighing procedures or by ocular estimation (U.S. Department of Agriculture, National Range and Pasture

Handbook, 2003). Where clipping-and-weighing procedures were done, two GPS coordinates were recorded – one at each end of an approximately 175 m transect (transects were actually 200 paces long, and so varied based on terrain and data-collector). The midpoint of these transects were used as evaluation data locations. One GPS coordinate was recorded at locations where ocular estimation was done, but personnel walked over areas averaging 4 hectares while estimating annual production (T. Staggs, Range Conservationist for the BLM, pers. comm., 2008). For these reasons, there was less confidence that species recorded were actually within 30 m of evaluation data GPS coordinates. Woody species that made up more than 5% of the total dry weight and herbaceous species that made up more than 2% of total dry weight were considered common and coded as "1" for evaluation purposes. Samples containing any non-native seeded species were eliminated from BLM data. Field observations indicated that the presence of non-native species was associated with lack of species diversity or abundance even if some native species remained.

The second source of evaluation data consisted of samples collected at 392 locations by a former graduate student for a passerine study at Utah State University (USU), Lindsay Brown (Brown, 2007). Unlike the BLM data, this data had high spatial accuracy. GPS coordinates were collected with a high quality GPS, and data were collected with a 10 m radius of coordinates. Locations were spread throughout the county, but many samples were spatially close to each other. USU samples closer than 90 m from another sample were not used for accuracy assessment. Fig. 3.6 shows the distribution of both BLM and USU sampling locations.

USU sampling involved the collection of both shrub height and foliar cover. Shrub species that were recorded in both height and foliar cover measurements were considered to be common. Herbaceous cover was estimated within the area of three Daubenmire frames (20 x 50 cm). In our opinion, estimating herbaceous cover only within three Daubenmire frames could have resulted in some species being missed. Therefore, herbaceous species that were identified in any of the Daubenmire frames were considered to be common.

Several USU samples were eliminated because the locations contained non-native seeded species; as with the BLM data, if any seeded species were present, the sample was not used for model evaluation. Another issue with these data was that no native grasses, and sometimes no native shrubs, were recorded at some locations. In this case, those samples were not used for evaluation purposes for either shrubs (if no native shrubs) or grasses (if no native grasses). This decision was made because it could not be ascertained whether these species were present but had been missed or whether they were truly absent or not common. Sagebrush species at several sampling locations were not identified to subspecies; those samples were not used to evaluate sagebrush models.

If a species was not found to be common (1-coded) in any samples in the combined evaluation dataset, that species model was not evaluated with independent data. Of the 38 species or species groups that had been modeled, 28 of them had at least one 1-coded sample in the evaluation data. Table 3.10 shows the number of 0- and 1-coded training records, USU, BLM, and USU+BLM records, and independent-data-estimated accuracy statistics are shown in Table 3.11.

2.4.2. Model evaluation using bootstrapped data

Because of the issues with the independent evaluation dataset mentioned above, it was decided that models should also be evaluated using some type of internal cross-validation procedure. The procedure we used was one where the original dataset was re-sampled (bootstrapped), while allowing records to be sampled more than once (sampled with replacement), with the number of bootstrap samples drawn being equal to the number of samples in the original dataset. This method, on average, draws 63.2% of the data for model parameter estimation, while leaving 36.8% for accuracy assessment (Steyerberg et al., 2001). Drawing approximately 63% of the data while leaving 37% for evaluation was very suitable for this dataset because some species had few 1-coded samples.

Some species, though, had so few 1-coded samples that the R software would frequently fail during accuracy assessment procedures; this occurred when 1-coded samples had not been drawn for either fitting or validation. For this reason it was decided that species with ten or less 1-coded samples would be evaluated slightly differently. In these cases, half of the dataset was randomly drawn *without replacement* for re-computing model variable coefficients; models were then evaluated using the remaining half of the data (50/50 accuracy assessment procedure). The bootstrap accuracy assessment procedure or alternative 50/50 procedure was repeated 100 times for each of the 38 species (or species groups). Estimated accuracy statistics were computed after each iteration and then averaged; results are shown in Table 3.12.

2.4.3. Comparison between independent data-estimated and bootstrap-estimated accuracies

Average bootstrap-sample-derived accuracy estimates were compared to those computed using independent (USU+BLM) data (Table 3.13). To determine whether bootstrap-estimated accuracy statistics were correlated to independent-data-estimated accuracy statistics, correlations were reviewed using the R software (Fig. 3.7 & Table 3.14).

2.5. Comparing three models using bootstrap-estimated accuracy statistics

Despite the fact that the BLM and USU data were not collected with this modeling effort in mind and had potential (spatial or correct classification) accuracy issues, we wanted to see if adding these data to our original training dataset would improve model accuracies. To this end, two models were created in addition to the original model fit using just the original training data (O/O fit) model. The first additional model used the original training data plus all of the independent USU+BLM evaluation data used for model evaluation (i.e. all of the data) to select model parameters and compute model coefficients (A/A fit). The second additional model used the same model parameters as the original model, but model coefficients were computed using all of the data (O/A fit). Based on the correlation between bootstrap-estimated and independent-data-estimated accuracy statistics (Table 3.13), it was decided that using bootstrap-estimated accuracy statistics would be a reasonable way to compare the three model fits.

For each of the three model fits, the bootstrap accuracy assessment (or alternative 50/50 accuracy assessment) procedure was repeated 100 times, and the average

sensitivity, specificity, S+S/2, and CCR computed. The best training datasets and models were chosen by determining which models had the best two out of following three accuracy estimates: sensitivity, S+S/2, and CCR (Table 3.14).

3. Results

The most important modeling variables (parameters), based on their frequency of selection for the original final models, include *slpd*, *tavg_pc1*, *ppt*, *ln_sca*, *sflux_pc1*, and *curv*. The *landsat_pc2* and *sflux_pc2* variables were also chosen for several species. Interaction variables were chosen less often than single variables.

Evaluation of model outputs generated from the original training data using the independent USU+BLM data indicated that models seemed to fit fairly well on average, despite the fact that there were known issues with the evaluation data. A summary of accuracy statistics computed using independent data is shown in Table 3.10. Sensitivity estimates were of particular interest because we wanted species to be predicted as common where they were common in evaluation data. Specificity, S+S/2, CCR were also computed for each species. For species that had at least ten 1-coded evaluation samples, the average sensitivity, specificity, S+S/2, and CCR accuracy estimates were 62.60%, 67.15%, 64.88%, and 68.28%, respectively. A few species models, such as those for *Leymus cinereus* (LECI4), *Pascopyrum smithii* (PASM), *Pseudoroegneria spicata* (PSSP6), and *Krascheninnikovia lanata* (KRLA2) gave poorer results based on independent data-estimated statistics. When there were fewer than ten 1-coded evaluation samples, estimated accuracy statistics were much more variable.

Accuracy statistics derived from internal cross validation using bootstrapped samples (or the alternative 50/50 procedure) were on average better than independent

data accuracy statistics, averaging 0.7344, 0.7722, and 0.8128 for sensitivity, S+S/2, and CCR respectively (Table 3.11). Comparisons of bootstrap-estimated accuracy statistics with accuracy statistics estimated using independent (USU+BLM) data indicated that the bootstrap-estimated accuracy statistics were generally higher (Table 3.13) and that there was a relationship between accuracy estimates obtained by the two methods (Table 3.14). The strongest relationship appeared to be between the specificity and CCR estimates, but sensitivity estimates were also significantly correlated when correlation tests were done using only species with more than ten 1-coded USU+BLM evaluation samples. The least correlated were S+S/2 estimates.

One of the most interesting results from these analyses was that data collected with this type of modeling effort in mind were better for selecting model parameters, but model coefficients computed using all available data often produced the most accurate models based on internal cross-validation procedures. Of the 28 species that had additional USU+BLM data to use for training, it was determined that 18 of them would best be modeled using the variables selected using just the original training data, but coefficients computed using the combined training data (O/A fit). Nine species were best modeled using just the original training data to determine variable sets and compute coefficients. Only one species (*Pascopyrum smithii*) had the best accuracy statistics when both model parameters and coefficients were computed using all of the available data (the A/A fit).

4. Discussion

Based on our results, logistic models predicting the spatial distribution of common species can be developed based on the relationship between species

distributions and abiotic attributes, including Landsat imagery. In general we were pleased with the results of these modeling efforts, especially considering the relatively small number of samples covering such a large area of diverse terrain. We would expect that model accuracies would improve if more data were available for modeling, particularly if data were collected with this type of modeling effort in mind – having both good spatial accuracy and a low chance of missing species. We were very not surprised by the modeling variables (parameters) chosen most frequently to produce the original final models, but it is interesting to note that the *elev* variable was chosen less frequently for models than the more directly resource-gradient-affecting slope, temperature, and precipitation variables. The *landsat_pc2* variable was also more frequently selected than *landsat_pc1*. The second PC of the seven Landsat bands highlighted green vegetation, whereas the first PC highlighted surface reflectance or brightness. This indicates, not surprisingly, that an index of the amount of photosynthetic material is more important than an index of surface reflectance for prediction of vegetation distributions. After reviewing the frequency with which variables were selected (Table 3.8), we were glad that we had included both *sflux_pc1* and *sflux_pc2* as potential model variables; *sflux_pc2* appeared to be an important model variable for several species.

There are a number of concerns when undertaking such a large sampling and modeling effort beyond obtaining a sufficient number of samples. One primary concern is that species may no longer be distributed across the landscape in a natural manner, or may be missing from some locations altogether. We might hypothesize that the poor sensitivity estimates of species such as *Leymus cinereus* (basin wildrye), *Artemisia tridentata* ssp. *tridentata* (basin big sagebrush), *Krascheninnikovia lanata* (winterfat) are

because these species have been greatly impacted by historic and/or current livestock grazing, or by human alteration of the landscape for agriculture. Low CCR estimates for *Pascopyrum smithii* (western wheatgrass), on the other hand, might be caused by the increase in dominance of this species on rangelands that, prior to European settlement, may have had a more diverse grass component. Model accuracies may also reflect differences in species' characteristics such as dispersal ability or longevity (Ozinga et al., 2005).

The fact that the USU+BLM data were not collected with this modeling effort in mind makes it a very good test of the original models, although estimated accuracies computed using these data may be a little lower than might be expected with a dataset collected with this type of modeling effort in mind (i.e. having both good spatial accuracy and low probability of missing species). We were fortunate to be able to additionally evaluate our 28 of our models with this independent data.

In the original training data, there were fewer than 20 1-coded samples for 16 of the species, and fewer than 10 for eight of the species. Models generated using fewer than 20 1-coded samples out of 366 might not be expected to be accurate. One could certainly question the logic of even trying to generate models using fewer than ten 1-coded samples. Species with few 1-coded samples were modeled anyway because they were felt to be important for ecological site description development and correlation. For assessment of the validity of these models we rely on the 50/50 procedure-produced accuracy statistics.

Comparisons of accuracy estimates of O/O, A/A, and O/A models were interesting, but it is a little more difficult to hypothesize the reason for the results.

Possibly the original data were better for choosing model variable sets because these data were collected with this modeling effort in mind. That still makes it unclear as to why using all of the data to generate parameter estimates appeared to increase estimated model accuracies based on bootstrap-estimated statistics.

5. References

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69.

Brown, L.J., 2007. A historic perspective: the response of breeding passerines to rangeland alteration in Rich County, Utah. Master's Thesis, Utah State University, Utah, 134 pp.

Crist, E.P., Cicone, R.C., 1984. A physically-basted transformation of Thematic Mapper data – the TM Tasseled Cap. IEEE Trans. Geosci. Remote. Sens. GS22, 256-263.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24(1), 38-49.

Guisan, A., Theurillat, J., 2000. Equilibrium modeling of alpine plant distribution: how far can we go? Phytocoenologia, 30, 353-384.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147-186.

Gleason, H.A., 1926. The individual concept of the plant association. In: McIntosh, R.P. (Ed.), Benchmark Papers in Ecology, vol. 6 – Phytosociology. Bull. Torrey Bot. Club, pp. 7-26.

Hall, C.A.S., Stanford, J.A., Hauer, F.R., 1992. The distribution and abundance of organisms as a consequence of energy balances along multiple environmental gradients. Oikos, 65, 377-390.

Hamilton, L.C., 1991. Regression with Graphics: A Second Course in Applied Statistics. Wadsworth, Inc., Belmont, CA, 363 pp.

Jensen, M.E., Haise, H.R., 1963. Estimating evapotranspiration from solar radiation. J. Irrig. Drainage Div. ASCE, 89, 15-41.

Kumar, L., Skidmore, A.K., Knowles, E., 1997. Modelling topographic variation in solar radiation in a GIS environment. Int. J. Geogr. Inf. Sci. 11, 475-497.

Lowry, J.L. Jr., Ramsey, R.D., Boykin, K., Bradford, D., Comer, P., Falzarano, S., Kepner, W., Kirby, J., Langs, L., Prior-Magee, J., Manis, G., O'Brien, L., Pohs, K., Rieth, W., Sajwaj, T., Schrader, S., Thomas, K.A., Schrupp, D., Schulz, K., Thompson, B., Wallace, C., Velasquez, C., Waller, E., Wolk, B. 2007. Mapping meso-scale land cover over very large geographic areas within a collaborative framework: A case study of the Southwest Regional Gap Analysis Project (SWReGAP). Remote Sens. Environ. 108, 59-73.

Ozinga, W.A., Schaminée, J.H.J., Bekker, R.M., Bonn, S., Poschlod, P., Tackenberg, O., Bakker, J., van Groenendael, J.M., 2005. Predictability of plant species composition from environmental conditions is constrained by dispersal limitation. Oikos, 108:555-561.

Park, S.J., McSweeney, K., Lowery, B., 2001. Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma 103, 249-272

Shultz, L. M., Monograph of *Artemisia* subgenus *Tridentatae*. (Asteraceae: Anthemideae). Systematic Botany Monographs (in press).

Steyerberg, E.W., Harrel, F.E. Jr., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y., Habbema, J.D.F. 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. 54, 774-781.

Tarboton, D.G. 1997. A new method for the determination of flow directions and contributing areas in grid digital elevation models. Water Resour. Res. 33, 309-319.

U.S. Department of Agriculture, Natural Resources Conservation Service, 2003. National Range and Pasture Handbook, title 190-VI-NRPH, Revision 1. [Online] Available: http://www.glti.nrcs.usda.gov/technical/publications/nrph.html.

U.S. Department of Agriculture, Natural Resources Conservation Service, 2007. National Soil Survey Handbook, title 430-VI. [Online] Available: http://soils.usda.gov/technical/handbook/

Walker, B., Kinzig, A., Langridge, J. 1999. Plant attribute diversity, resilience, and ecosystem function: The nature and significance of dominant and minor species. Ecosyst. 2, 95-113.

Zimmermann, N.E., Tools for analyzing, summarizing, and mapping of biophysical variables. [Online] Available: http://www.wsl.ch/staff/niklaus.zimmermann/progs.html

Table 3.1. Eigenvalues and factor loadings for the first two principal components of the 12 solar flux grids.

Eigenvalues:	29283448.29	1099951.808
% of Variance:	96.19%	3.61%
Cumulative % Variance:	96.19%	99.80%

Eigenvectors:

	PC1	PC2
sflux_01	0.31929	-0.22884
sflux_02	0.36597	-0.16349
sflux_03	0.35448	-0.00879
sflux_04	0.27247	0.2154
sflux_05	0.16503	0.41908
sflux_06	0.10505	0.5155
sflux_07	0.13389	0.46764
sflux_08	0.23145	0.29045
sflux_09	0.32916	0.06431
sflux_10	0.36555	-0.12046
sflux_11	0.33972	-0.22046
sflux_12	0.31159	-0.25055

Table 3.2. Eigenvalues and factor loadings for the first five principal components of the 12 average temperature grids. Only the first 3 components were used in modeling.

Eigenvalues:	30451.97	23707.84	955.70	527.93	398.15
% of Variance:	54.01%	42.05%	1.70%	0.94%	0.71%
Cumulative % Variance:	54.01%	96.06%	97.75%	98.69%	99.40%

Eigenvectors

	PC1	PC2	PC3	PC4	PC5
tavg_01	-0.02128	-0.57163	-0.15289	-0.05119	-0.35598
tavg_02	-0.04395	-0.54117	0.1832	0.48784	-0.08419
tavg_03	0.16372	-0.23846	-0.45202	0.43081	0.62378
tavg_04	0.40082	0.10206	-0.34717	0.03108	-0.07039
tavg_05	0.42695	0.1359	-0.00176	0.03616	-0.18094
tavg_06	0.41774	0.10211	-0.02927	0.03858	-0.04159
tavg_07	0.3499	-0.01369	0.49842	0.07589	0.2179
tavg_08	0.28549	-0.09507	0.48118	-0.06073	0.26719
tavg_09	0.28493	-0.09513	0.11345	0.02507	-0.16312
tavg_10	0.31029	-0.08533	-0.09549	0.13003	-0.48774
tavg_11	0.26458	-0.11749	-0.32952	-0.4763	0.1827
tavg_12	0.04424	-0.49481	0.08263	-0.56231	0.14463

Table 3.3. Correlation matrix for several climatic variables considered for use in modeling. Based on this table, *tavg_ann*, *tmax_07*, *tmin_01*, *tmax-tmin*, *aridity_idx*, and *seasonality_idx* variables were not used in logistic models because they were highly correlated with other variables. Bold type indicates correlation coefficients >0.50, underlining indicates correlation coefficients >0.85.

	elev					
ppt_ann	**0.899**	*ppt_ann*				
tavg_ann	-0.427	-0.331	*tavg_ann*			
tavg_pc1	**-0.801**	**-0.687**	**0.865**	*tavg_pc1*		
tavg_pc2	-0.538	**-0.531**	**-0.502**	0.000	*tavg_pc2*	
tavg_pc3	0.054	0.191	-0.010	0.000	0.000	*tavg_pc3*
tmax_07	**-0.967**	**-0.906**	0.495	**0.844**	0.467	0.010
tmin_01	0.724	**0.728**	0.271	-0.232	**-0.942**	0.179
tmax-tmin	**-0.928**	**-0.896**	0.128	**0.593**	**0.769**	-0.091
aridity_idx	**0.893**	**0.984**	-0.403	**-0.740**	-0.477	0.144
seasonality_idx	**0.874**	**0.983**	-0.238	**-0.617**	**-0.595**	0.151

Table 3.4. Eigenvalues and factor loadings for the first four principal components of the seven Landsat TM bands, including the thermal band. The Landsat image extended to a 1 km buffered boundary of Rich County. The first three bands were used in modeling; subsequent bands were discarded.

Eigenvalues:	1687.0626	173.5872	46.7517	11.47214
% of Variance:	87.41%	8.99%	2.42%	0.59%
Cumulative % Variance:	87.41%	96.40%	98.82%	99.42%

Eigenvectors:

Landsat band				
1	0.0848	-0.1536	0.4048	0.4426
2	0.1629	-0.0604	0.4890	0.2548
3	0.2746	-0.0787	0.5426	-0.0594
4	0.4075	0.8230	0.0089	-0.1691
5	0.6556	-0.1595	-0.4931	0.5211
6	0.4994	-0.4658	0.0093	-0.6594
7	0.2135	0.2166	0.2437	-0.0286

Table 3.5. Correlation matrix for 13 logistic modeling variables. Bold font indicates correlation coefficients > 0.50.

	elev											
slpd	0.461	*slpd*										
curv	0.050	0.049	*curv*									
ln_sca	-0.190	-0.178	-0.456	*ln_sca*								
ppt	**0.893**	0.452	0.007	-0.069	*ppt*							
tavg_pc1	**-0.785**	-0.329	-0.007	0.176	**-0.671**	*tavg_pc1*						
tavg_pc2	**-0.540**	-0.309	-0.006	0.044	**-0.544**	-0.005	*tavg_pc2*					
tavg_pc3	0.079	0.171	0.003	0.063	0.184	0.027	-0.024	*tavg_pc3*				
sflux_pc1	-0.054	-0.194	0.023	-0.004	-0.072	0.034	0.068	-0.034	*sflux_pc1*			
sflux_pc2	-0.321	**-0.894**	-0.034	0.108	-0.348	0.219	0.211	-0.167	-0.024	*sflux_pc2*		
landsat_pc1	-0.003	-0.047	0.044	-0.210	-0.180	-0.203	0.209	-0.319	0.465	-0.001	*landsat_pc1*	
landsat_pc2	0.229	0.067	-0.008	0.045	0.300	-0.240	0.003	0.027	-0.046	-0.082	-0.008	*landsat_pc2*
landsat_pc3	-0.250	-0.076	0.115	-0.041	-0.281	0.206	0.171	-0.093	0.081	0.014	0.001	0.004

Table 3.6. Correlation matrix for 15 logistic modeling interaction variables. Bold font indicates correlation coefficients > 0.50.

	slpd:curv													
slpd:ln_sca	-0.326	*slpd:ln_sca*												
slpd:ppt	0.037	-0.063	*slpd:ppt*											
slpd:tavg_pc1	-0.017	0.166	**-0.680**	*slpd:tavg_pc1*										
slpd:sflux_pc1	0.007	0.010	-0.053	0.019	*slpd:sflux_pc1*									
curv:ln_sca	-0.095	0.050	0.058	-0.062	-0.006	*curv:ln_sca*								
curv:ppt	0.325	-0.126	0.050	-0.029	0.001	0.167	*curv:ppt*							
curv:tavg_pc1	-0.210	0.069	-0.032	0.040	-0.001	-0.103	**-0.721**	*curv:tavg_pc1*						
curv:sflux_pc1	-0.204	0.061	-0.006	0.005	0.080	-0.094	-0.091	0.034	*curv:sflux_pc1*					
ln_sca:ppt	-0.097	**0.544**	-0.090	0.129	0.002	-0.183	**-0.50**2	0.348	0.042	*ln_sca:ppt*				
ln_sca:tavg_pc1	0.030	**-0.511**	0.052	-0.289	0.004	0.111	0.222	-0.312	-0.010	**-0.599**	*ln_sca:tavg_pc1*			
ln_sca:sflux_pc1	0.085	-0.216	0.027	-0.038	-0.290	0.171	0.076	-0.022	**-0.587**	-0.161	0.100	*ln_sca:sflux_pc1*		
ppt:tavg_pc1	-0.002	0.020	-0.354	0.431	0.012	-0.005	-0.015	0.017	-0.003	0.059	-0.096	-0.005	*ppt:tavg_pc1*	
ppt:sflux_pc1	0.003	0.004	-0.154	0.110	0.480	0.013	0.028	-0.020	0.024	-0.018	0.007	-0.075	0.025	*ppt:sflux_pc1*
tavg_pc1:sflux_pc1	-0.002	-0.022	0.135	-0.195	-0.298	0.001	-0.020	0.026	-0.017	0.000	0.025	0.069	-0.049	**-0.761**

Table 3.7. Variables and interactions to be considered for use in logistic regression models.

elev	elevation
slpd	slope in degrees
curv	curvature
ln_sca	natural logarithm of specific catchment area
ppt	annual average total precipitation
tavg_pc1	principal component 1 derived from 12 monthly average temperature grids
tavg_pc2	principal component 2 derived from 12 monthly average temperature grids
tavg_pc3	principal component 3 derived from 12 monthly average temperature grids
sflux_pc1	principal component 1 derived from 12 solar flux grids
sflux_pc2	principal component 2 derived from 12 solar flux grids
landsat_pc1	principal component 1 derived from seven Landsat bands
landsat_pc2	principal component 2 derived from seven Landsat bands
landsat_pc3	principal component 3 derived from seven Landsat bands
slpd:curv	
slpd:ln_sca	
slpd:ppt	
slpd:tavg_pc1	
slpd:sflux_pc1	
curv:ln_sca	
curv:ppt	
curv:tavg_pc1	
curv:sflux_pc1	
ln_sca:ppt	
ln_sca:tavg_pc1	
ln_sca:sflux_pc1	
ppt:tavg_pc1	
ppt:sflux_pc1	
tavg_pc1:sflux_pc1	

Table 3.8. List of species/species groups modeled. The N=1 column indicates the number of sample locations at which the species was common.

Plant Code	Common Name	Scientific Name	N=1
ACHY	Indian ricegrass	*Achnatherum hymenoides* (Roem. & Schult.) Barkworth	44
ACLE9	Letterman's needlegrass	*Achnatherum lettermanii* (Vasey) Barkworth	89
ACNE9	Columbia needlegrass	*Achnatherum nelsonii* (Scribn.) Barkworth	12
AMELA	serviceberry	*Amelanchier* Medik.	36
ARAR8	little sagebrush	*Artemisia arbuscula* Nutt.	30
ARNO4	black sagebrush	*Artemisia nova* A. Nelson	26
ARTRB	Bonnevillensis big sagebrush	*Artemisia tridentata* spp. "*bonnevillensis*" (UNOFFICIAL, Shultz, 2009)	20
ARTRS2	snowfield sagebrush	*Artemisia tridentata* Nutt. ssp. *Spiciformis* (Osterh.) Kartesz & Gandhi	4
ARTRT	basin big sagebrush	*Artemisia tridentata* Nutt. ssp. *tridentata*	13
ARTRV	mountain big sagebrush	*Artemisia tridentata* Nutt. ssp. *vaseyana* (Rydb.) Beetle	59
ARTRW8	Wyoming big sagebrush	*Artemisia tridentata* Nutt. ssp. *wyomingensis* Beetle & Young	140
ATGA	Gardner's saltbush	*Atriplex gardneri* (Moq.) D. Dietr.	18
BRMA4	mountain brome	*Bromus marginatus* Nees ex Steud.	14
CAGE2	Geyer's sedge	*Carex geyeri* Boott	11
CAREX_W	wetland sedge + others		20
CARO5	Ross' sedge	*Carex rossii* Boott	43
CELE3	curl-leaf mtn mahogany	*Cercocarpus ledifolius* Nutt.	3
CONIF	any subalpine conifers		32
ELEL5	squirreltail	*Elymus elymoides* (Raf.) Swezey	51
ELTR7	slender wheatgrass	*Elymus trachycaulus* (Link) Gould ex Shinners	30
FEOV	sheep fescue	*Festuca ovina* L.	7
HECOC8	needle-and-thread	*Hesperostipa comata* (Trin. & Rupr.) Barkworth ssp. *comata*	53
JUOS	Utah juniper	*Juniperus osteosperma* (Torr.) Little	7
KOMA	prairie Junegrass	*Koeleria macrantha* (Ledeb.) Schult.	30
KRLA2	winterfat	*Krascheninnikovia lanata* (Pursh) A. Meeuse & Smit	7
LECI4	basin wildrye	*Leymus cinereus* (Scribn. & Merr.) A. Löve	10
LEKI2	spike fescue	*Leucopoa kingii* (S. Watson) W.A. Weber	9
PASM	western wheatgrass	*Pascopyrum smithii* (Rydb.) A. Löve	77
POFE	muttongrass	*Poa fendleriana* (Steud.) Vasey	82
POPR	Kentucky bluegrass	*Poa pratensis* L.	21
POSE	Sandberg bluegrass	*Poa secunda* J. Presl	180
POTR5	aspen	*Populus tremuloides* Michx.	13
PRVI	chokecherry	*Prunus virginiana* L.	7
PSSP6	bluebunch wheatgrass	*Pseudoroegneria spicata* (Pursh) A. Löve	84
PUTR2	bitterbrush	*Purshia tridentata* (Pursh) DC.	26
SALIX	willow	*Salix* L.	6
SAVE4	greasewood	*Sarcobatus vermiculatus* (Hook.) Torr.	11
SYOR2	mountain snowberry	*Symphoricarpos oreophilus* A. Gray	66

Table 3.9. The number of times each modeling parameter was selected for final models and the average P-value for those variables in the final fitted logistic regression models. Bold font indicates a non-interaction variable.

Herbaceous Species			Woody Species			All 39 Species Models		
Variable	Times Selected	Avg Pr>|z|	Variable	Times Selected	Avg Pr(>|z|)	Variable	Times Selected	Avg Pr(>|z|)
ln_sca	**19**	**0.2070**	***slpd***	**20**	**0.3220**	***slpd***	**39**	**0.2776**
slpd	**19**	**0.2309**	***tavg_pc1***	**19**	**0.3050**	***tavg_pc1***	**38**	**0.2793**
ppt	**19**	**0.2414**	***sflux_pc1***	**19**	**0.3171**	***ppt***	**37**	**0.2189**
tavg_pc1	**19**	**0.2536**	***curv***	**19**	**0.3609**	***ln_sca***	**37**	**0.2408**
sflux_pc1	**18**	**0.2456**	***ppt***	**18**	**0.1951**	***sflux_pc1***	**37**	**0.2823**
curv	**18**	**0.3221**	***ln_sca***	**18**	**0.2765**	***curv***	**37**	**0.3420**
landsat_pc2	**12**	**0.0381**	***sflux_pc2***	**11**	**0.0856**	***landsat_pc2***	**22**	**0.0922**
ppt:tavg_pc1	**11**	**0.0654**	*slpd:ln_sca*	10	0.1446	***sflux_pc2***	**21**	**0.1017**
sflux_pc2	**10**	**0.1193**	***landsat_pc2***	**10**	**0.1571**	*ppt:tavg_pc1*	17	0.0496
tavg_pc2	**8**	**0.0339**	*curv:ln_sca*	8	0.1454	*slpd:ln_sca*	16	0.2605
tavg_pc3	**8**	**0.0426**	*ln_sca:sflux_pc1*	8	0.1472	***landsat_pc3***	**15**	**0.0441**
ln_sca:tavg_pc1	8	0.0651	***landsat_pc3***	**7**	**0.0078**	***landsat_pc1***	**15**	**0.1059**
landsat_pc3	**8**	**0.0759**	***landsat_pc1***	**7**	**0.0919**	*curv:tavg_pc1*	15	0.1946
landsat_pc1	**8**	**0.1181**	***elev***	**7**	**0.2362**	***tavg_pc2***	**14**	**0.0259**
curv:tavg_pc1	8	0.1245	*curv:tavg_pc1*	7	0.2747	*ln_sca:sflux_pc1*	14	0.1650
ppt:sflux_pc1	7	0.1348	***tavg_pc2***	**6**	**0.0153**	*curv:ln_sca*	14	0.1749
slpd:ppt	7	0.2098	*ppt:tavg_pc1*	6	0.0207	***elev***	**13**	**0.1498**
elev	**6**	**0.0489**	*tavg_pc1:sflux_pc1*	6	0.0896	*tavg_pc1:sflux_pc1*	12	0.1267
slpd:sflux_pc1	6	0.0856	*curv:sflux_pc1*	6	0.1151	*slpd:ppt*	12	0.1334
tavg_pc1:sflux_pc1	6	0.1637	*slpd:sflux_pc1*	6	0.2271	*slpd:sflux_pc1*	12	0.1563
ln_sca:sflux_pc1	6	0.1887	*slpd:ppt*	5	0.0264	*ppt:sflux_pc1*	12	0.2046
curv:ln_sca	6	0.2143	*ln_sca:ppt*	5	0.2340	***tavg_pc3***	**11**	**0.0501**
slpd:ln_sca	6	0.4537	*ppt:sflux_pc1*	5	0.3024	*ln_sca:tavg_pc1*	10	0.0584
ln_sca:ppt	5	0.0460	*slpd:tavg_pc1*	4	0.0767	*ln_sca:ppt*	10	0.1400
slpd:tavg_pc1	5	0.1439	***tavg_pc3***	**3**	**0.0700**	*curv:sflux_pc1*	10	0.1439
curv:sflux_pc1	4	0.1871	*slpd:curv*	3	0.3528	*slpd:tavg_pc1*	9	0.1140
slpd:curv	3	0.0698	*ln_sca:tavg_pc1*	2	0.0316	*slpd:curv*	6	0.2113
curv:ppt	2	0.0005	*curv:ppt*	2	0.0978	*curv:ppt*	4	0.0491

Table 3.10. Summary of 0- and 1-coded training samples and 1-coded USU and BLM evaluation samples. Note: plant codes ARTRT and UARTRT refer to *Artemisia tridentata* ssp. *tridentata*, bottomland-occurring and upland-occurring, respectively.

Plant	Plant	Training Samples		USU's	BLM's	USU+BLM Samples	
Form	Code	1	0	1-coded	1-coded	1	0
Trees	CONIF	32	334	---	---	---	---
	JUOS	7	359	---	---	---	---
	POTR5	13	353	0	3	3	64
Shrubs	AMELA	36	308	15	11	26	320
	ARAR8	30	314	1	9	10	338
	ARNO4	26	318	26	9	35	313
	ARTRB	20	324	---	---	---	---
	ARTRS2	4	340	---	---	---	---
	ARTRT	13	331	4	4	8	340
	UARTRT	16	328	---	---	---	---
	ARTRV	59	285	0	13	13	54
	ARTRW8	140	204	0	28	28	39
	ATGA	18	326	---	---	---	---
	CELE3	3	341	0	1	1	66
	KRLA2	7	337	3	10	13	333
	PRVI	7	337	0	2	2	65
	PUTR2	26	318	15	12	27	319
	SALIX	6	338	---	---	---	---
	SAVE4	11	333	6	0	6	273
	SYOR2	66	278	28	18	46	300
Grasses and Grass-Like	ACHY	44	282	25	16	41	249
	ACLE9	89	327	0	7	7	60
	ACNE9	12	314	1	5	6	284
	BRMA4	14	312	0	3	3	64
	CAGE2	11	315	---	---	---	---
	CAREX	20	306	---	---	---	---
	CARO5	43	283	---	---	---	---
	ELEL5	51	275	83	11	94	196
	ELTR7	30	296	2	0	2	221
	FESTU	7	319	0	6	6	61
	HECOC8	53	273	37	27	64	226
	KOMA	30	296	11	14	25	265
	LECI4	10	316	3	8	11	279
	LEKI2	9	317	---	---	---	---
	PASM	77	249	55	41	96	194
	POFE	82	244	5	42	47	243
	POPR	21	305	0	10	10	57
	POSE	180	146	51	57	108	182
	PSSP6	84	242	79	32	111	179

Table 3.11. Results of evaluation of original models with independent (USU+BLM) data. Bold type indicates that those species had at least ten 1-coded evaluation samples. [Sens. = sensitivity, Spec. = specificity; S+S/2 = (sensitivity + specificity) / 2; CCR = overall correct classification rate]

Plant	Training Samples		Evaluation Samples		Independent Data Estimated Accuracy			
Code	1	0	1	0	Sens.	Spec.	S+S/2	CCR
ACHY	**44**	**284**	**41**	**249**	**58.5%**	**67.1%**	**62.8%**	**65.9%**
ACLE9	89	239	7	60	85.7%	31.7%	58.7%	37.3%
ACNE9	12	316	6	284	50.0%	82.4%	66.2%	81.7%
AMELA	**36**	**310**	**26**	**320**	**73.1%**	**75.9%**	**74.5%**	**75.7%**
ARAR8	**30**	**316**	**10**	**338**	**40.0%**	**74.6%**	**57.3%**	**73.6%**
ARNO4	**26**	**320**	**35**	**313**	**45.7%**	**81.5%**	**63.6%**	**77.9%**
ARTRT	13	333	8	340	12.5%	64.7%	38.6%	63.5%
ARTRV	**59**	**287**	**13**	**54**	**76.9%**	**70.4%**	**73.6%**	**71.6%**
ARTRW8	**140**	**206**	**28**	**39**	**75.0%**	**79.5%**	**77.2%**	**77.6%**
BRMA4	14	314	3	64	0.0%	100.0%	50.0%	95.5%
CELE3	3	343	1	66	100.0%	87.9%	93.9%	88.1%
ELEL5	**51**	**277**	**94**	**196**	**52.1%**	**66.3%**	**59.2%**	**61.7%**
ELTR7	30	298	2	221	100.0%	33.9%	67.0%	34.5%
FESTU	7	321	6	61	100.0%	83.6%	91.8%	85.1%
HECOC8	**53**	**275**	**64**	**226**	**68.8%**	**64.2%**	**66.5%**	**65.2%**
KOMA	**30**	**298**	**25**	**265**	**92.0%**	**58.5%**	**75.2%**	**61.4%**
KRLA2	**7**	**339**	**13**	**333**	**38.5%**	**70.3%**	**54.4%**	**69.1%**
LECI4	**10**	**318**	**11**	**279**	**27.3%**	**91.0%**	**59.2%**	**88.6%**
PASM	**77**	**251**	**96**	**194**	**79.2%**	**26.3%**	**52.7%**	**43.8%**
POFE	**82**	**246**	**47**	**243**	**51.1%**	**61.3%**	**56.2%**	**59.7%**
POPR	**21**	**307**	**10**	**57**	**60.0%**	**77.2%**	**68.6%**	**74.6%**
POSE	**180**	**148**	**108**	**182**	**91.7%**	**28.6%**	**60.1%**	**52.1%**
POTR5	13	355	3	64	66.7%	100.0%	83.3%	98.5%
PRVI	7	339	2	65	0.0%	98.5%	49.2%	95.5%
PSSP6	**84**	**244**	**111**	**179**	**47.7%**	**62.6%**	**55.2%**	**56.9%**
PUTR2	**26**	**320**	**27**	**319**	**66.7%**	**79.0%**	**72.8%**	**78.0%**
SAVE4	11	335	6	273	83.3%	91.2%	87.3%	91.0%
SYOR2	**66**	**280**	**46**	**300**	**82.6%**	**74.7%**	**78.6%**	**75.7%**
Computed using species with at least ten 1-coded evaluation samples				**Mean:**	**62.6%**	**67.2%**	**64.9%**	**68.3%**
				Max:	**92.0%**	**91.0%**	**78.6%**	**88.6%**
				Min:	**27.3%**	**26.3%**	**52.7%**	**43.8%**
Computed using species with at least one 1-coded evaluation sample				Mean:	61.6%	70.8%	66.2%	71.4%
				Max:	100.0%	100.0%	93.9%	98.5%
				Min:	0.0%	26.3%	38.6%	34.5%

Table 3.12. Accuracy statistics for the original models computed using 100 iterations of the bootstrap cross-validation procedure or alternative 50/50 procedure. Bold font indicates that statistics were computed using the bootstrap cross-validation procedure and that there were at least 10 1-coded samples in the dataset for that species. [Sens. = sensitivity, Spec. = specificity; S+S/2 = (sensitivity + specificity) / 2; CCR = overall correct classification rate]

Plant	Bootstrap Cross-Validation Estimated			
Code	Sens.	Spec.	S+S/2	CCR
ACHY	**88.0%**	**80.1%**	**84.0%**	**81.0%**
ACLE9	**81.6%**	**74.5%**	**78.1%**	**76.3%**
ACNE9	**37.5%**	**94.8%**	**66.2%**	**92.7%**
AMELA	**81.8%**	**77.2%**	**79.5%**	**77.6%**
ARAR8	**72.2%**	**81.5%**	**76.8%**	**80.7%**
ARNO4	**78.3%**	**77.7%**	**78.0%**	**77.6%**
ARTRB	**69.8%**	**79.1%**	**74.5%**	**78.5%**
ARTRS2	20.9%	98.7%	59.8%	97.5%
ARTRT	**50.1%**	**91.3%**	**70.7%**	**89.5%**
ARTRV	**85.8%**	**80.8%**	**83.3%**	**81.6%**
ARTRW8	**82.4%**	**88.5%**	**85.4%**	**85.9%**
ATGA	**59.0%**	**88.4%**	**73.7%**	**86.7%**
BRMA4	**75.3%**	**93.0%**	**84.2%**	**92.1%**
CAGE2	**79.8%**	**95.2%**	**87.5%**	**94.6%**
CAREX_W	**78.5%**	**97.3%**	**87.9%**	**96.1%**
CARO5	**84.5%**	**67.8%**	**76.2%**	**70.0%**
CELE3	24.0%	98.1%	61.0%	97.2%
CONIF	**78.7%**	**97.7%**	**88.2%**	**96.1%**
ELEL5	**80.4%**	**52.6%**	**66.5%**	**56.9%**
ELTR7	**63.6%**	**69.5%**	**66.5%**	**69.0%**
FESTU	32.4%	95.9%	64.1%	94.4%
HECOC8	**83.0%**	**77.8%**	**80.4%**	**78.6%**
JUOS	55.2%	96.7%	75.9%	95.8%
KOMA	**78.2%**	**59.9%**	**69.0%**	**61.7%**
KRLA2	44.2%	95.1%	69.7%	93.9%
LECI4	**33.6%**	**90.6%**	**62.1%**	**88.8%**
LEKI2	27.1%	94.5%	60.8%	92.5%
PASM	**87.9%**	**44.2%**	**66.0%**	**54.5%**
POFE	**85.4%**	**75.8%**	**80.6%**	**78.1%**
POPR	**41.9%**	**93.6%**	**67.7%**	**90.1%**
POSE	**80.3%**	**76.0%**	**78.2%**	**78.4%**
POTR5	**78.9%**	**96.2%**	**87.5%**	**95.5%**
PRVI	43.1%	98.5%	70.8%	97.2%
PSSP6	**81.1%**	**75.1%**	**78.1%**	**76.6%**
PUTR2	**79.4%**	**86.0%**	**82.7%**	**85.3%**
SALIX	63.9%	98.6%	81.2%	97.9%
SAVE4	**67.4%**	**93.0%**	**80.2%**	**92.1%**
SYOR2	**78.9%**	**74.8%**	**76.8%**	**75.5%**

Statistics computed using only species with at least 10 1-coded samples.

	Mean:	Max:	Min:
Sens.	**73.4%**	**88.0%**	**33.6%**
Spec.	**81.0%**	**97.7%**	**44.2%**
S+S/2	**77.2%**	**88.2%**	**62.1%**
CCR	**81.3%**	**96.1%**	**54.5%**

Statistics computed using all species.

	Mean:	Max:	Min:
Sens.	66.2%	88.0%	20.9%
Spec.	84.4%	98.7%	44.2%
S+S/2	75.3%	88.2%	59.8%
CCR	84.3%	97.9%	54.5%

Table 3.13. Comparison of accuracy estimates produced fro the bootstrap cross-validation procedure and evaluation using independent (USU+BLM) data. Bold type indicates that there were more than ten 1-coded USU+BLM evaluation samples for that species. [Sens. = sensitivity, Spec. = specificity; S+S/2 = (sensitivity + specificity) / 2; CCR = overall correct classification rate]

Plant	Bootstrap Cross-Validation Estimated				Independent Data Estimated				Difference (Bootstrapped - Independent)			
Code	Sens.	Spec.	S+S/2	CCR	Sens.	Spec.	S+S/2	CCR	Sens.	Spec.	S+S/2	CCR
ACHY	**88.0%**	**80.1%**	**84.0%**	**81.0%**	**58.5%**	**67.1%**	**62.8%**	**65.9%**	**29.5%**	**13.0%**	**21.2%**	**15.2%**
ACLE9	81.6%	74.5%	78.1%	76.3%	85.7%	31.7%	58.7%	37.3%	-4.1%	42.8%	19.4%	39.0%
ACNE9	37.5%	94.8%	66.2%	92.7%	50.0%	82.4%	66.2%	81.7%	-12.5%	12.4%	0.0%	11.0%
AMELA	**81.8%**	**77.2%**	**79.5%**	**77.6%**	**73.1%**	**75.9%**	**74.5%**	**75.7%**	**8.7%**	**1.3%**	**5.0%**	**1.9%**
ARAR8	**72.2%**	**81.5%**	**76.8%**	**80.7%**	**40.0%**	**74.6%**	**57.3%**	**73.6%**	**32.2%**	**7.0%**	**19.6%**	**7.1%**
ARNO4	**78.3%**	**77.7%**	**78.0%**	**77.6%**	**45.7%**	**81.5%**	**63.6%**	**77.9%**	**32.6%**	**-3.8%**	**14.4%**	**-0.3%**
ARTRT	50.1%	91.3%	70.7%	89.5%	12.5%	64.7%	38.6%	63.5%	37.6%	26.6%	32.1%	26.0%
ARTRV	**85.8%**	**80.8%**	**83.3%**	**81.6%**	**76.9%**	**70.4%**	**73.7%**	**71.6%**	**8.9%**	**10.5%**	**9.7%**	**10.0%**
ARTRW8	**82.4%**	**88.5%**	**85.4%**	**85.9%**	**75.0%**	**79.5%**	**77.2%**	**77.6%**	**7.4%**	**9.0%**	**8.2%**	**8.3%**
BRMA4	75.3%	93.0%	84.2%	92.1%	0.0%	100.0%	50.0%	95.5%	75.3%	-7.0%	34.2%	-3.4%
CELE3	24.0%	98.1%	61.0%	97.2%	100.0%	87.9%	93.9%	88.1%	-76.0%	10.2%	-32.9%	9.1%
ELEL5	**80.4%**	**52.6%**	**66.5%**	**56.9%**	**52.1%**	**66.3%**	**59.2%**	**61.7%**	**28.2%**	**-13.7%**	**7.3%**	**-4.8%**
ELTR7	63.6%	69.5%	66.5%	69.0%	100.0%	33.9%	67.0%	34.5%	-36.4%	35.5%	-0.4%	34.5%
FESTU	32.4%	95.9%	64.1%	94.4%	100.0%	83.6%	91.8%	85.1%	-67.6%	12.3%	-27.7%	9.3%
HECOC8	**83.0%**	**77.8%**	**80.4%**	**78.6%**	**68.8%**	**64.2%**	**66.5%**	**65.2%**	**14.3%**	**13.6%**	**14.0%**	**13.4%**
KOMA	**78.2%**	**59.9%**	**69.0%**	**61.7%**	**92.0%**	**58.5%**	**75.3%**	**61.4%**	**-13.8%**	**1.4%**	**-6.2%**	**0.3%**
KRLA2	**44.2%**	**95.1%**	**69.7%**	**93.9%**	**38.5%**	**70.3%**	**54.4%**	**69.1%**	**5.8%**	**24.8%**	**15.3%**	**24.8%**
LECI4	**33.6%**	**90.6%**	**62.1%**	**88.8%**	**27.3%**	**91.0%**	**59.2%**	**88.6%**	**6.3%**	**-0.5%**	**2.9%**	**0.2%**
PASM	**87.9%**	**44.2%**	**66.0%**	**54.5%**	**79.2%**	**26.3%**	**52.7%**	**43.8%**	**8.7%**	**17.9%**	**13.3%**	**10.7%**
POFE	**85.4%**	**75.8%**	**80.6%**	**78.1%**	**51.1%**	**61.3%**	**56.2%**	**59.7%**	**34.3%**	**14.4%**	**24.4%**	**18.4%**
POPR	**41.9%**	**93.6%**	**67.7%**	**90.1%**	**60.0%**	**77.2%**	**68.6%**	**74.6%**	**-18.1%**	**16.4%**	**-0.9%**	**15.5%**
POSE	**80.3%**	**76.0%**	**78.2%**	**78.4%**	**91.7%**	**28.6%**	**60.1%**	**52.1%**	**-11.3%**	**47.4%**	**18.0%**	**26.3%**
POTR5	78.9%	96.2%	87.5%	95.5%	66.7%	100.0%	83.3%	98.5%	12.2%	-3.8%	4.2%	-3.0%
PRVI	43.1%	98.5%	70.8%	97.2%	0.0%	98.5%	49.2%	95.5%	43.1%	0.1%	21.6%	1.7%
PSSP6	**81.1%**	**75.1%**	**78.1%**	**76.6%**	**47.8%**	**62.6%**	**55.2%**	**56.9%**	**33.4%**	**12.5%**	**23.0%**	**19.7%**
PUTR2	**79.4%**	**86.0%**	**82.7%**	**85.3%**	**66.7%**	**79.0%**	**72.8%**	**78.0%**	**12.7%**	**7.0%**	**9.9%**	**7.3%**
SAVE4	67.4%	93.0%	80.2%	92.1%	83.3%	91.2%	87.3%	91.0%	-16.0%	1.8%	-7.1%	1.0%
SYOR2	**78.9%**	**74.8%**	**76.8%**	**75.5%**	**82.6%**	**74.7%**	**78.6%**	**75.7%**	**-3.7%**	**0.1%**	**-1.8%**	**-0.2%**
Average calculated using models with at least ten 1-coded external validation samples								**Mean:**	**12.0%**	**9.9%**	**11.0%**	**9.7%**
Average calculated using models at least one 1-coded external validation samples								Mean:	11.0%	11.0%	8.6%	10.7%

Table 3.14. Correlation between bootstrap cross-validation and independent data (USU+BLM) accuracy assessment statistics. P-values are for a 0.95 threshold 1-tailed test.

	All Pairs		Pairs with at least 10 1-coded evaluation samples	
Statistic	**corr (r)**	**P**	**corr (r)**	**P**
Sensitivity	0.162	0.205	0.550	0.009
Specificity	0.717	< 0.001	0.638	0.002
(Sens + Spec) / 2	0.021	0.458	0.390	0.055
CCR	0.729	< 0.001	0.643	0.002

Table 3.15. Comparison of the average of 100 bootstrap cross-validation (or alternative 50/50 procedure) accuracy estimates for three different dataset/model combinations. O/O estimates are for the model variables selected using the original training data and coefficients computed using the original training dataset. A/A estimates are for the model variables selected using the original training data plus incorporated USU+BLM data (i.e. all of the data), and coefficients computed using this same dataset. O/A estimates are for model variables selected using the original training data, but with coefficients computed using all of the data. Empty entries indicate that there were no more data available to add to the original training data. The 'Best Model' column indicates the dataset/model combination where two of the three following statistics were greatest: sensitivity, (sensitivity + specificity) / 2, and Correct Classification Rate. Bold type indicates the dataset/model combination with the greatest estimated accuracy assessment value for the statistic indicated at the top of the columns for the section. Table continues on the next page.

	Sensitivity			Specificity			Sensitivity + Specificity / 2			Correct Classification Rate			Best
Species	O/O	A/A	O/A	O/O	A/A	O/A	O/O	A/A	O/A	O/O	A/A	O/A	Model
ACHY	**88.0%**	79.6%	79.6%	**80.1%**	75.6%	75.9%	**84.0%**	77.6%	77.8%	**81.1%**	76.1%	76.4%	O/O
ACLE9	**81.6%**	77.7%	79.6%	74.5%	74.1%	**80.5%**	78.1%	75.9%	**80.1%**	76.3%	74.9%	**80.3%**	O/A
ACNE9	37.5%	61.1%	**95.0%**	**94.8%**	88.7%	91.6%	66.2%	74.9%	**93.3%**	**92.7%**	87.8%	91.7%	O/A
AMELA	81.8%	**89.5%**	89.3%	77.2%	74.6%	**79.6%**	79.5%	82.0%	**84.4%**	77.6%	75.9%	**80.4%**	O/A
ARAR8	72.2%	68.5%	**73.8%**	**81.5%**	72.3%	75.4%	**76.9%**	70.4%	74.6%	**80.7%**	72.1%	75.3%	O/O
ARNO4	**78.3%**	70.1%	72.7%	77.7%	73.8%	**79.0%**	**78.0%**	72.0%	75.8%	77.6%	73.5%	**78.4%**	O/O
ARTRB	**69.8%**	---	---	**79.1%**	---	---	**74.5%**	---	---	**78.5%**	---	---	O/O
ARTRS2	**20.9%**	---	---	**98.7%**	---	---	**59.8%**	---	---	**97.5%**	---	---	O/O
ARTRT	50.1%	66.3%	**81.0%**	**91.3%**	75.4%	78.4%	70.7%	70.9%	**79.7%**	**89.5%**	75.0%	78.4%	O/A
ARTRV	85.8%	85.0%	**88.8%**	80.8%	78.3%	**85.7%**	83.3%	81.7%	**87.3%**	81.7%	79.5%	**86.3%**	O/A
ARTRW8	82.4%	82.3%	**83.4%**	88.5%	85.6%	**88.6%**	85.4%	83.9%	**86.0%**	**86.0%**	84.2%	86.5%	O/A
ATGA	**59.0%**	---	---	**88.4%**	---	---	**73.7%**	---	---	**86.7%**	---	---	O/O
BRMA4	75.3%	63.7%	**94.0%**	93.0%	**98.0%**	95.0%	84.2%	80.9%	**94.5%**	92.1%	**96.5%**	94.9%	O/A
CAGE2	**79.8%**	---	---	**95.2%**	---	---	**87.5%**	---	---	**94.6%**	---	---	O/O
CARO5	**84.5%**	---	---	**67.8%**	---	---	**76.2%**	---	---	**70.1%**	---	---	O/O
CAREX_W	**78.5%**	---	---	**97.3%**	---	---	**87.9%**	---	---	**96.1%**	---	---	O/O
CELE3	24.0%	17.9%	**100.0%**	98.1%	98.6%	**98.7%**	61.1%	58.3%	**99.3%**	97.2%	97.6%	**98.7%**	O/A
CONIF	**78.7%**	---	---	**97.7%**	---	---	**88.2%**	---	---	**96.1%**	---	---	O/O
ELEL5	80.4%	80.6%	**80.6%**	52.6%	47.8%	**53.8%**	66.5%	64.2%	**67.2%**	56.9%	55.8%	**60.1%**	O/A
ELTR7	63.6%	63.2%	**78.2%**	**69.5%**	66.1%	64.9%	66.5%	64.6%	**71.5%**	**69.0%**	65.9%	65.6%	O/A
FESTU	32.4%	45.8%	**100.0%**	**95.9%**	95.6%	93.0%	64.1%	70.7%	**96.5%**	**94.4%**	93.8%	93.3%	O/A

Table 3.15. Continuation from previous page.

	Sensitivity			Specificity			Sensitivity + Specificity / 2			Correct Classification Rate			Best
Species	O/O	A/A	O/A	O/O	A/A	O/A	O/O	A/A	O/A	O/O	A/A	O/A	model
HECOC8	**83.0%**	78.6%	77.6%	**77.8%**	67.2%	71.7%	**80.4%**	72.9%	74.7%	**78.7%**	69.3%	72.8%	O/O
JUOS	**55.2%**	---	---	**96.7%**	---	---	**75.9%**	---	---	**95.8%**	---	---	O/O
KOMA	78.2%	83.7%	**92.5%**	59.9%	63.6%	**63.9%**	69.1%	73.7%	**78.2%**	61.6%	65.4%	**66.4%**	O/A
KRLA2	44.2%	72.1%	**90.8%**	**95.1%**	76.0%	62.0%	69.7%	74.0%	**76.4%**	**93.9%**	75.8%	62.8%	O/A
LECI4	33.6%	74.2%	**96.8%**	**90.6%**	81.7%	83.0%	62.1%	78.0%	**89.9%**	**88.8%**	81.4%	83.4%	O/A
LEKI2	**27.2%**	---	---	**94.5%**	---	---	**60.8%**	---	---	**92.5%**	---	---	O/O
PASM	**87.9%**	58.4%	62.3%	44.2%	**71.1%**	68.7%	66.0%	64.7%	**65.5%**	54.5%	**67.5%**	67.0%	A/A
POFE	**85.4%**	74.9%	81.1%	**75.8%**	66.2%	68.9%	**80.6%**	70.5%	75.0%	**78.1%**	68.0%	71.5%	O/O
POPR	41.9%	75.8%	**95.1%**	**93.6%**	86.4%	88.5%	67.7%	81.1%	**91.8%**	**90.1%**	85.5%	89.0%	O/A
POSE	80.3%	**77.9%**	77.7%	**76.0%**	66.7%	68.4%	**78.2%**	72.3%	73.0%	**78.4%**	71.9%	72.8%	O/O
POTR5	78.9%	67.6%	**100.0%**	96.2%	97.2%	**97.2%**	87.5%	82.4%	**98.6%**	95.5%	96.0%	**97.4%**	O/A
PRVI	43.1%	48.1%	**99.7%**	**98.5%**	98.3%	91.5%	70.8%	73.2%	**95.6%**	**97.2%**	97.2%	91.7%	O/A
PSSP6	**81.1%**	70.9%	79.9%	**75.1%**	68.9%	63.9%	**78.1%**	69.9%	71.9%	**76.7%**	69.5%	69.0%	O/O
PUTR2	79.4%	82.5%	**87.3%**	**86.0%**	74.5%	78.0%	**82.7%**	78.5%	82.6%	**85.3%**	75.1%	78.7%	O/O
SALIX	**63.9%**	---	---	**98.6%**	---	---	**81.2%**	---	---	**97.9%**	---	---	O/O
SAVE4	**67.4%**	33.1%	0.0%	93.0%	97.4%	**99.5%**	**80.2%**	65.3%	49.7%	92.1%	95.6%	**96.6%**	O/O
SYOR2	78.9%	83.1%	**85.7%**	74.8%	77.2%	**80.6%**	76.8%	80.1%	**83.1%**	75.5%	78.1%	**81.4%**	O/A

Best Model Mean Sens: 79.5%
Best Model Mean Spec: 83.8%
Best Model Mean S+S/2: 81.7%
Best Model Mean CCR: 83.9%

Figures

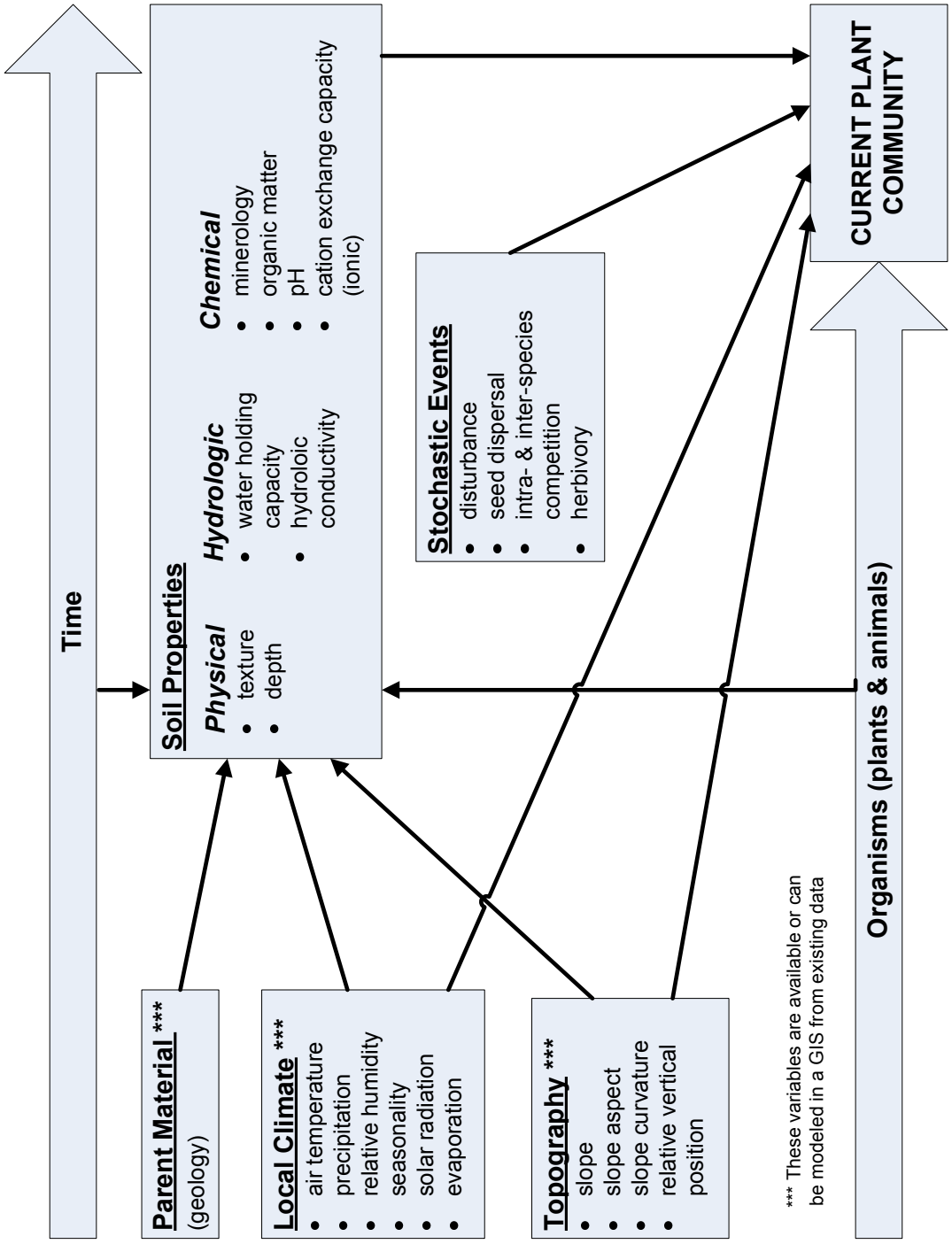

Fig. 3.1. Conceptual model showing variables which drive plant species distribution in semi-arid environments.

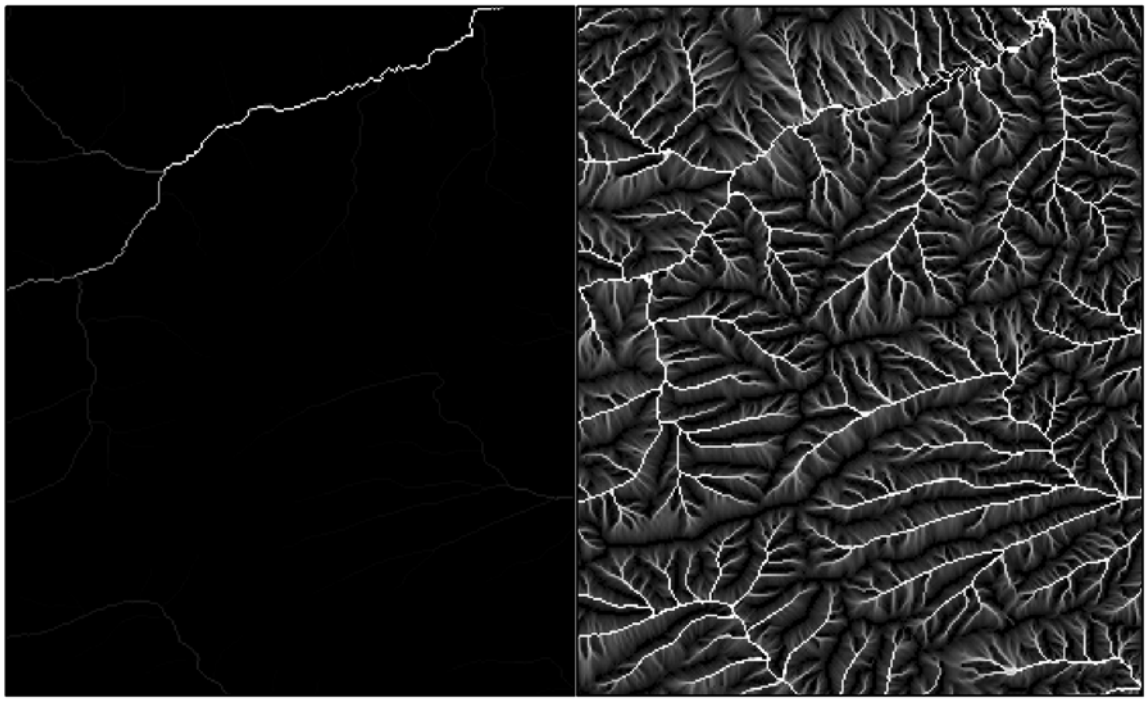

Fig. 3.2. Illustration of specific catchment (*sca*) raw values compared to natural logarithm transformed *sca* values (*ln_sca*). The raw *sca* grid on the left had a minimum of 30, maximum of 18,697,444, mean of 157,973, and standard deviation of 1,656,326. The *ln_sca* grid on the right had a minimum of 3.40, maximum of 16.74, mean of 5.27, and standard deviation of 2.

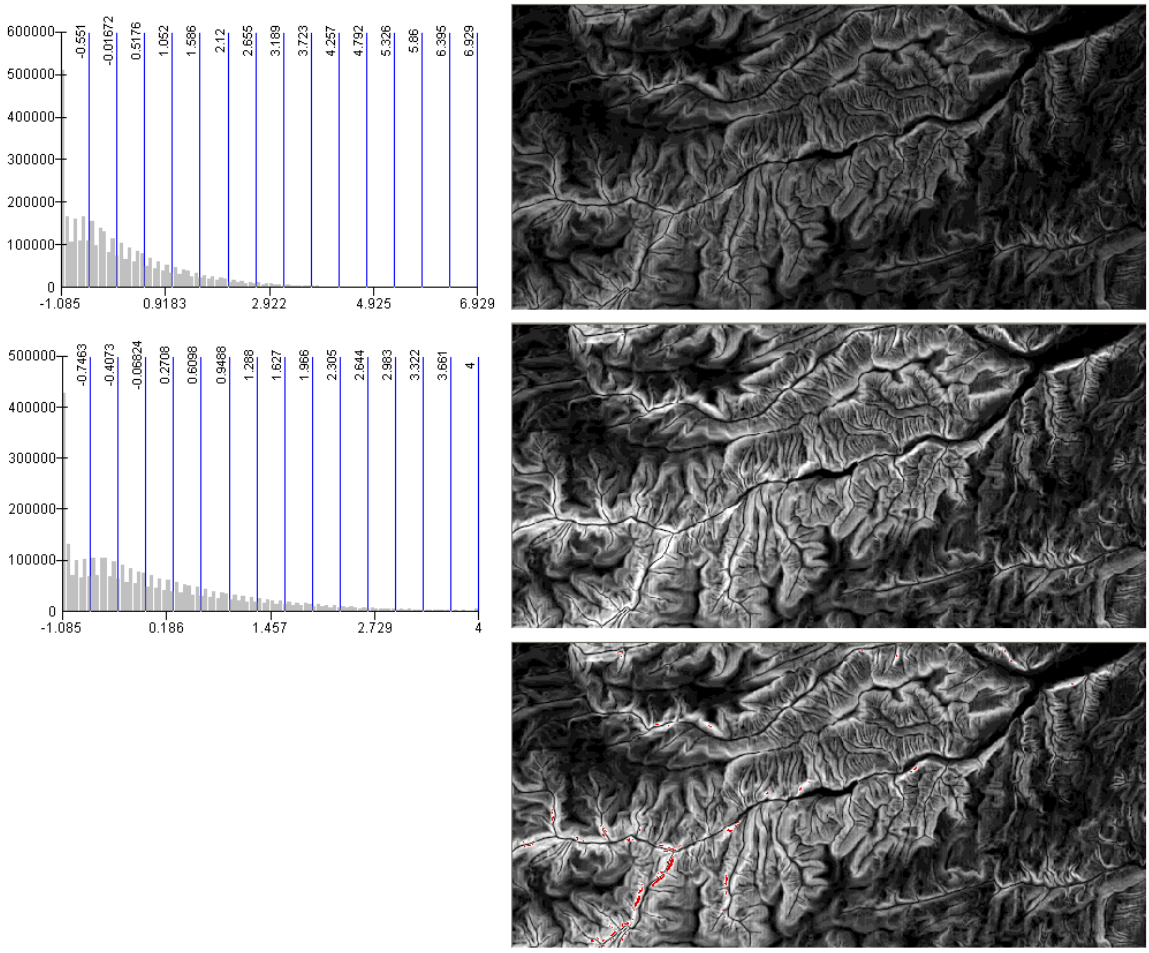

Fig. 3.3. Illustration of why variables were cut off at four standard deviations above or below the mean. The variable shown here is slope, which initially had values almost seven standard deviations above the mean. Red areas in the lower map identify pixels that have values more than four standard deviations above the mean; these accounted for about 0.1% of slope values. Grayscale of maps is in 15 equal intervals.

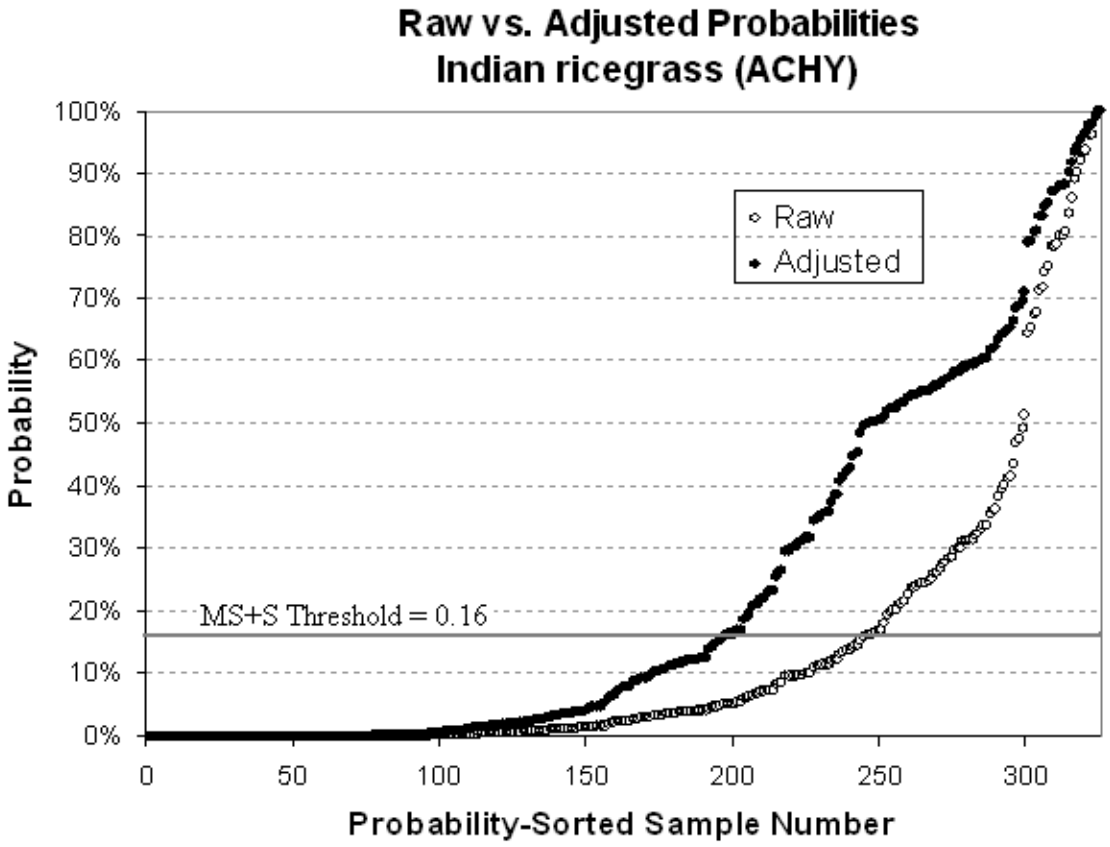

Fig. 3.4. Example showing how logistic model probability-value outputs were adjusted to normalize thresholds between common and not-common to 0.5 while maintaining probability values between 0 and 1.

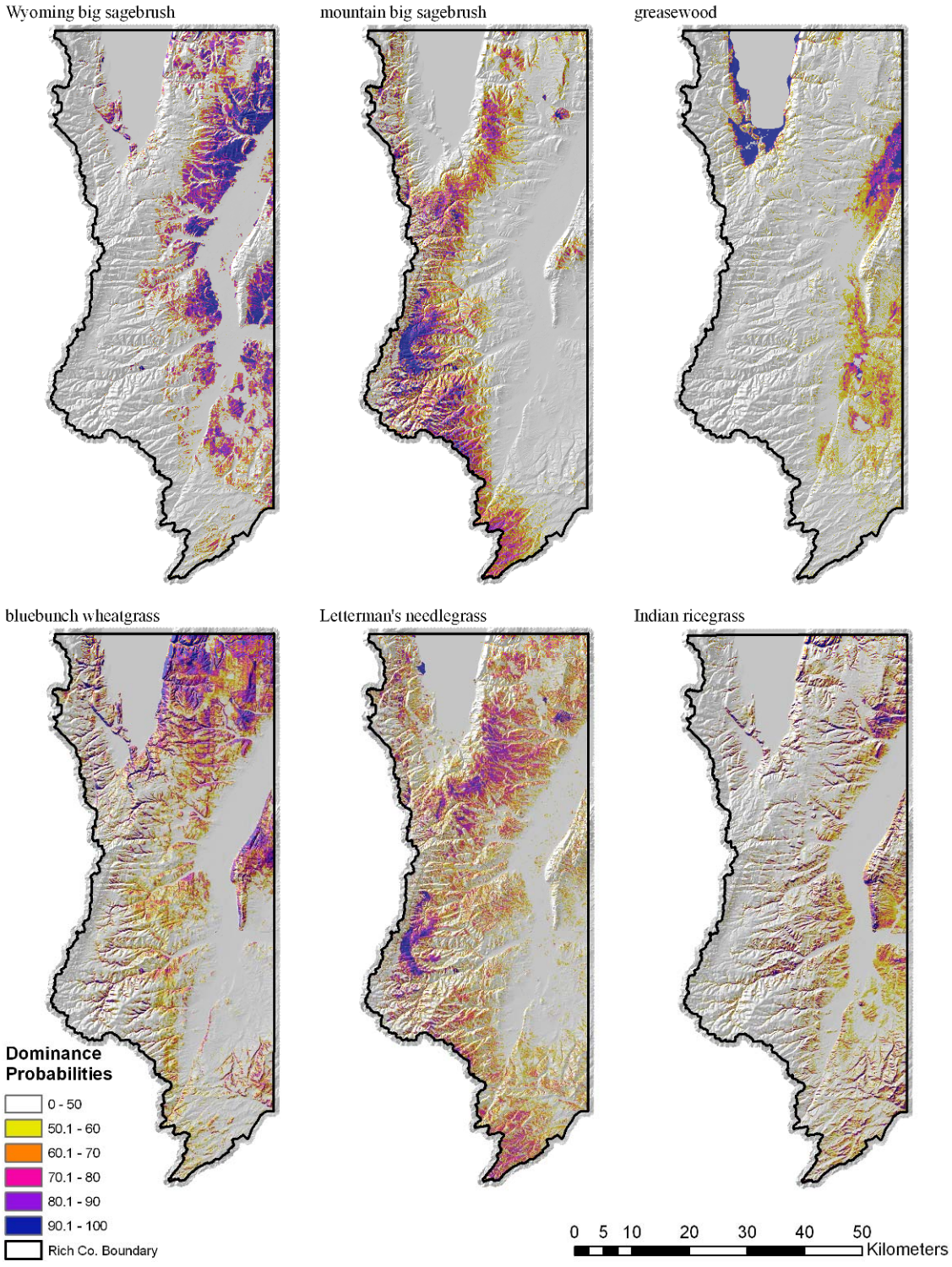

Fig. 3.5. Examples of threshold-standardized logistic regression model outputs. Probabilities are shown as percent values rather than decimal values.

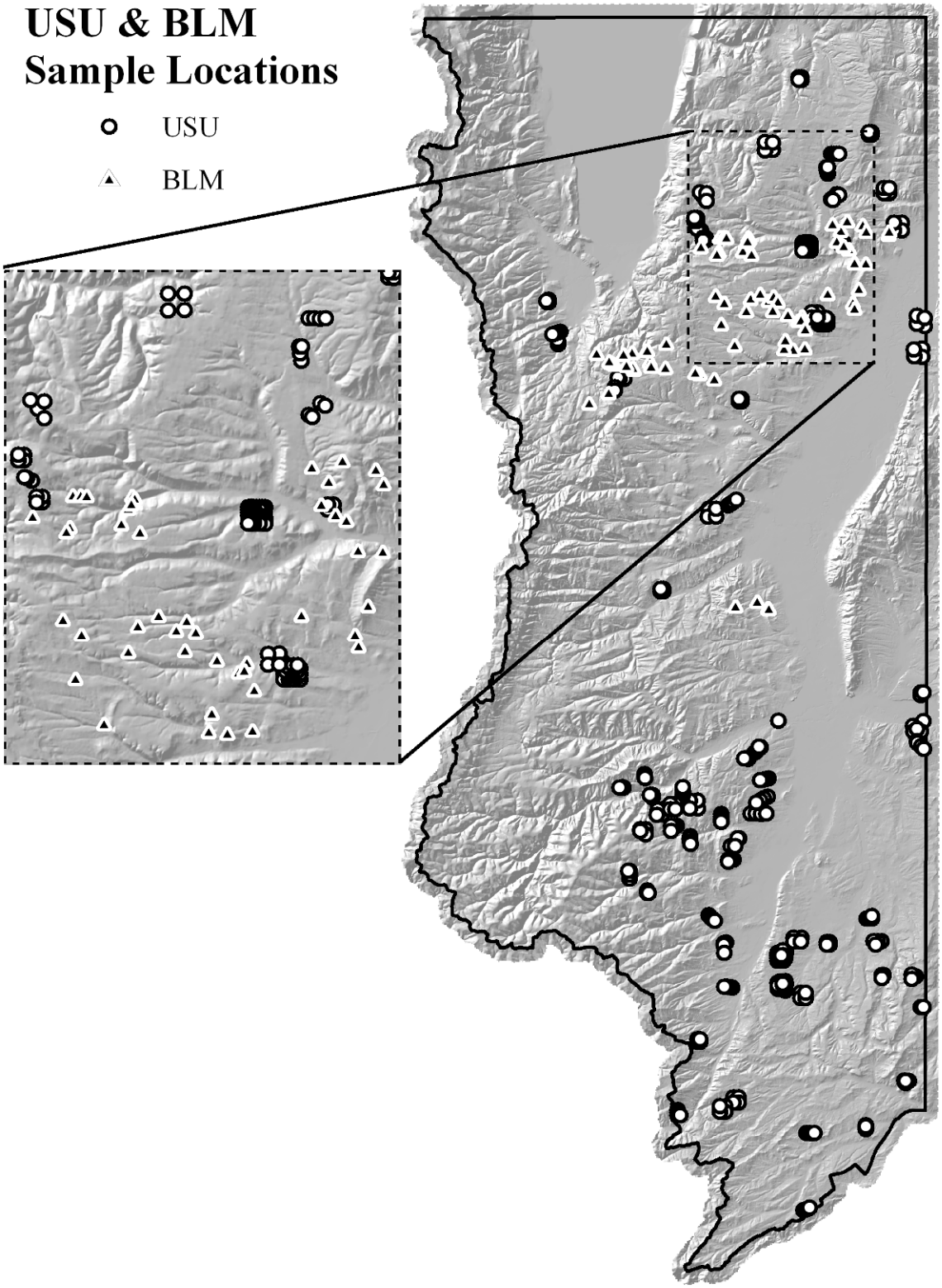

Fig. 3.6. Distribution BLM and USU evaluation sample locations.

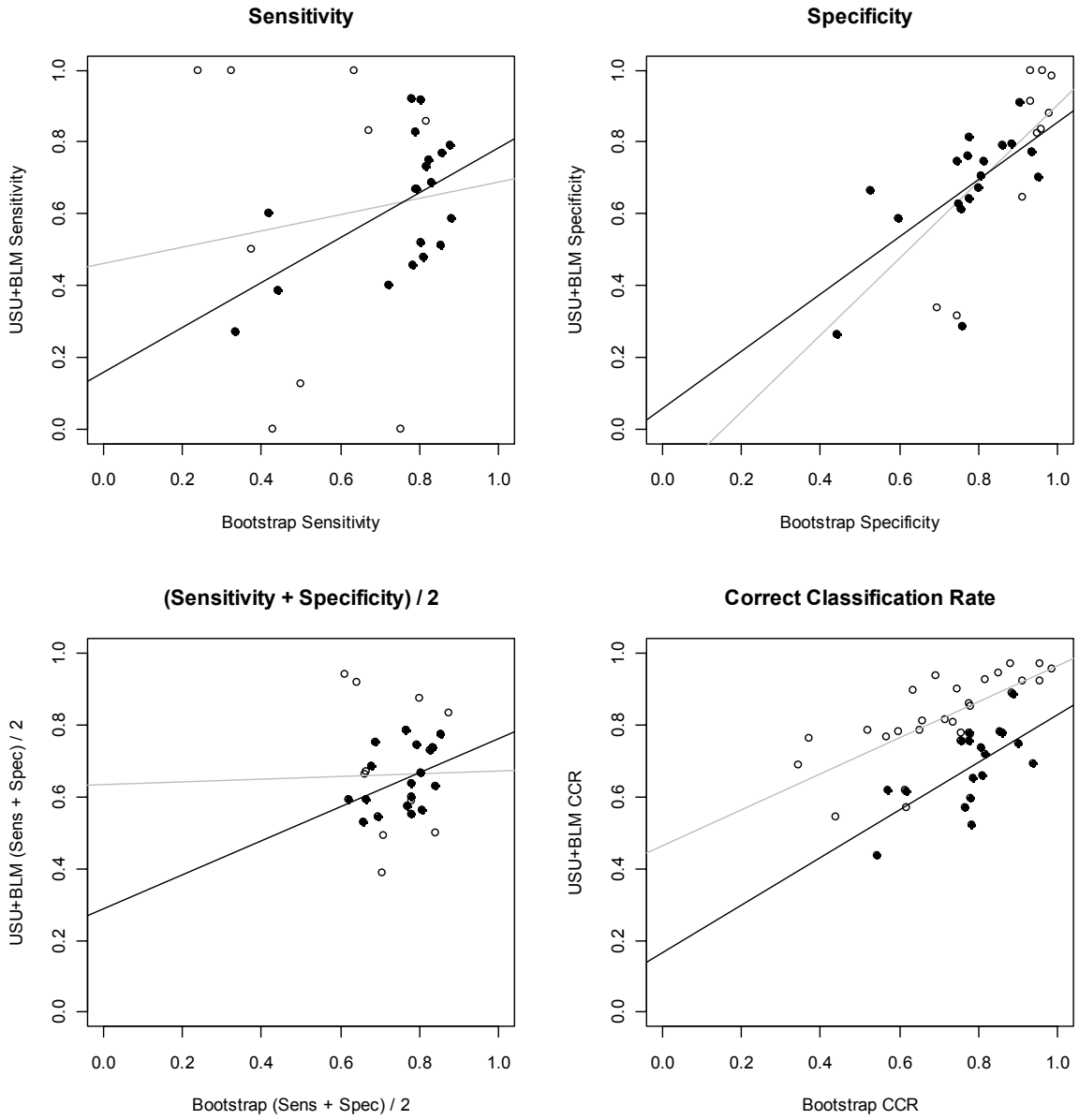

Fig. 3.7. Correlation between bootstrap cross-validation and independent data (USU+BLM) accuracy assessment statistics. Solid dots and indicate species with at least 10 1-coded samples; circles represent species with less than 10 1-coded samples. Black lines are the regression lines through only the solid dots; grey lines are regression lines thorough all of the points.

CHAPTER 4

DISCUSSION

The goal of this project was to evaluate ESD to soil map unit correlations in Rich County through the use of potential common species distribution models. The work outlined in this thesis significantly contributed towards this goal. Published soil survey information can be used in conjunction with species distribution models to assist with the correlation processes.

Field Sampling

The use of the ISODATA algorithm to cluster abiotic attributes to drive the sampling effort proved to be an effective way to stratify the landscape. Even so, we felt like there is room for improvement with this methodology. Even with the method's apparent weaknesses, it provided an excellent guide by which to canvas the county and gain insight into the relationship between abiotic factors and the spatial distribution of plant communities.

Potential Common Species Modeling

A significant amount of time was spent on this part of the project, and we do not believe that there was much more we could have done in terms of modeling methodology or variable inclusion that would have improved the accuracy of the resulting models. We felt that using multiple subsets of the dataset to choose final model variables was a good strategy, especially once we figured out how sensitive stepwise model selection procedures were to the datasets being used. We were also very happy with using the

"maximum sensitivity + specificity" threshold criterion to classify species as potentially common or not common.

Model accuracies may have been improved by other factors, such as by including more sample data, specifically data that had been collected for this type of modeling effort. Data collected by ourselves or for the Southwest Re-GAP Analysis project in 2001 only included 366 samples for the entire county. Fortunately we were able to augment the data by including data from the BLM and the parallel USU study; this provided up to 692 samples for some species. Still, some species were common in only a few locations; their models should be considered highly suspect.

Another factor that may have affected accuracies, especially in areas with little topographic relief, might have been the resolution of Landsat and DEM data. The DEM and Landsat data were at 30 m resolution; the climatic data (originally at 800 m resolution) was resampled to the same resolution. If finer resolution data are used, model accuracies may improve.

Classifying species in the sample data as either common or not common may have had an effect on models as well. As was noted, some known decreaser grass species were very sparse in many locations. We made the assumption that these species had the potential to be common in these locations even though their foliar cover was less than 1%. It is possible that we were mistaken in some of these cases; that the species would not have been common even with less grazing pressure. This would certainly affect the clarity of the data and the accuracy of the models. The best solution to this problem would have been to have spent enough time in the field to find a sufficient number of less altered communities to sample so that samples from marginally diverse communities

could have been eliminated from the data. Even with more time, though, it would be unlikely that a large number of less impacted plant communities would have been found in some lower-elevation areas that have been mostly modified for agricultural production of hay and/or livestock.